UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO CONHECIMENTO

Rodrigo Bittencourt Cabral

CONCEPÇÃO, IMPLEMENTAÇÃO E VALIDAÇÃO DE UM ENFOQUE PARA INTEGRAÇÃO E RECUPERAÇÃO DE CONHECIMENTO DISTRIBUÍDO EM BASES DE DADOS HETEROGÊNEAS

Dissertação submetida ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Engenharia do Conhecimento
Orientador: Prof. Dr. rer. nat. Aldo v. Wangenheim
Co-orientador: Prof. Dr. José Leomar Todesco

Florianópolis

2010

Catalogação na fonte pela biblioteca da
Universidade Federal de Santa Catarina

C117c Cabral, Rodrigo Bittencourt
Concepção, implementação e validação de um enfoque para integração e recuperação de conhecimento distribuído em bases de dados heterogêneas [dissertação] / Rodrigo Bittencourt Cabral ; orientador, Aldo von Wangenheim. - Florianópolis, SC, 2010.
178 p.: il., grafs., tabs.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

Inclui referências

1. Engenharia e gestão do conhecimento. 2. Toxicologia. 3. Ontologia. 4. Semântica. 5. Sistemas de recuperação da informação. I. Wangenheim, Aldo v. (Aldo von). II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. III. Título.

CDU 659.2

Rodrigo Bittencourt Cabral

CONCEPÇÃO, IMPLEMENTAÇÃO E VALIDAÇÃO DE UM ENFOQUE PARA INTEGRAÇÃO E RECUPERAÇÃO DE CONHECIMENTO DISTRIBUÍDO EM BASES DE DADOS HETEROGÊNEAS

Esta Dissertação foi julgada adequada para obtenção do Título de "mestre", e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento

Florianópolis, 01 de outubro de 2010.

_____
Prof. Roberto Carlos dos Santos Pacheco, Dr.
Coordenador do Curso

Banca Examinadora:

_____
Prof., Dr. rer. nat. Aldo v. Wangenheim,
Orientador
Universidade Federal de Santa Catarina

_____
Prof., Dr. José Leomar Todesco,
Co-Orientador
Universidade Federal de Santa Catarina

_____
Prof., Dr. Mario Antônio Ribeiro Dantas,
Universidade Federal de Santa Catarina

_____
Profª, Drª. Francis Solange Vieira Tourinho,
Universidade Federal do Rio Grande do Norte

À minha esposa Fabiane, meus pais João e Loreni e meus avós Alberto e Ruth.

AGRADECIMENTOS

A Deus - detentor de todo o conhecimento e poder - pela vida, pela misericórdia, por Jesus e pela Salvação, e também por ter trazido-me até aqui.
À minha esposa Fabiane, pelo amor, companheirismo, cumplicidade, por ter compartilhado comigo as alegrias e por ter me consolado nas tristezas.
Aos professores Aldo e Tite, que me orientaram não só na realização deste trabalho, mas que me trouxeram exemplos e orientações para a vida.
À professora Christiane, que me trouxe várias idéias criativas para possibilitar o desenvolvimento deste trabalho.
À professora Marlene, Danielle e Marizete, que possibilitaram que o contexto do estudo de caso deste trabalho fosse explorado aprofundadamente.
Aos meus pais João Batista e Loreni, por terem enfrentado todas as dificuldades para me dar a criação maravilhosa que tive.
Aos meus avós Alberto e Ruth, por sempre terem despendido todo o apoio necessário com bondade, bom humor e otimismo.
Aos meus tios Roberto e Lucélia, que me apoiaram desde o começo nesta nova empreitada, e sempre acreditaram na possibilidade de sucesso.
Aos meus pastores/amigos Carlos e Raquel, que estiveram ao meu lado me dando apoio espiritual e emocional, me orientando nas situações adversas por que passei durante o trabalho, e compartilhando da felicidade nestes momentos.
Aos meus irmãos em Cristo Jean, Fernando, Jonathan e Faustino, pela colaboração e amizade que sempre prestaram a mim.
Aos meus colegas de trabalho Harley, Andrade, Cloves, Marcone, Savaris, Douglas, Andrei, Magnos, Coelho, Marcos, Marcus Vinicius, Pavezi, Ruby, Eros, Chris, Camile, Antonio, Coser e Cleidson, que sempre estiveram ao meu lado, otimistas, sempre me auxiliando quando preciso.

“A parte importante do progresso é o desejo por progresso.”
(Sêneca)

RESUMO

Com o crescimento da demanda e da composição de Bases de Conhecimento para os mais diversos fins e a sua disponibilização através da rede mundial de computadores, passou-se a observar a necessidade de organizar este conhecimento e também integrá-lo para possibilitar maior acessibilidade e facilidade na sua manutenção e utilização, devido à caracterização da disposição dispersa e o formato heterogêneo das referidas bases. Neste trabalho é proposto um sistema que efetua integração do conhecimento de bases de dados em contexto genérico, utilizando como estudo de caso o atendimento emergencial no CIT - Centro de Informações Toxicológicas de Santa Catarina - além de possibilitar a manutenção e manipulação deste artefato através do agrupamento de técnicas de recuperação de informação, aperfeiçoamento semântico, expansão de consulta, fonética em um único mecanismo. Foram avaliadas - através de uma revisão sistemática da literatura - as melhores opções disponibilizadas por estudos prévios em pesquisas realizadas nestas áreas a fim de encontrar a melhor combinação a ser utilizada no mecanismo, além da análise do produto final em um comparativo feito entre mecanismos previamente utilizados pelos profissionais no atendimento de urgência.

Palavras-chave: Toxicologia Clínica, Base de Conhecimento, Engenharia de Conhecimento, Ontologia, Mecanismo de Busca, Expansão de Consulta, Semântica.

ABSTRACT

With growth demand and composition of knowledge bases for different purposes and making them available through internet, it's possible to see the need to organize this knowledge and also integrate it to provide greater accessibility and ease maintenance and use, due to the characterization of dispersed persistence and format of such heterogeneous databases. This dissertation proposes a system that performs integration of knowledge databases in generic context, using as a case study of emergency care at CIT - Toxicological Information Center of Santa Catarina - besides facilitating the maintenance and manipulation of the artifact by grouping techniques of information retrieval, semantic processing, query expansion, phonetics in a single mechanism. Were evaluated - through a systematic literature review - the best options available in previous studies on research conducted in these areas to find the best combination to be used in the mechanism, besides the analysis of the final product in a comparison made between mechanisms previously used by professionals in emergency care.

Keywords: Clinical Toxicology, Knowledge Base, Knowledge Engineering, Ontology, Search Engine, Query Expansion, Semantics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANVISA	Agência Nacional de Vigilância Sanitária
API	*application Programming Interface*
CDSS	*clinical decision support system*
CIT	centro de Informações Toxicológicas
CTD	*Comparative Toxicogenomics Database*
DCB	Denominações Comuns Brasileiras
DeCS	Descritores em Ciências da Saúde
GQM	*goal query metrics*
HIS	*hospital information systems*
MESH	*Medical Subject Headings*
MVC	*model view controller*
HP	*Hewlett Packard®*
HSDB	*Hazardous Substances Data Bank*
JSON	*JavaScript Object Notation*
SPARQL	*SPARQL Protocol and RDF Query Language*
SQL	*Structured Query Language*
PACS	*Picture Archiving and Communication System*
P@10	*precision at ten*
PDF	*Portable Document Format*
PLN	Processamento de linguagem natura
RF	*relevance feedback*
PRF	*pseudo relevance feedback*
QE	*query expansion*
RDF	*resource description format*
RIS	*Radiology Information System*
SBC	sistema baseado em conhecimento
SA	*semantic annotations*
SI	*semantic improvement*
UMLS	*Unified Medical Language System*
XML	*eXtensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO

1.1. CONTEXTUALIZAÇÃO DO PROBLEMA

O Centro de Toxicologia de Santa Catarina (CIT/SC) desenvolve seu trabalho integrando as três grandes áreas de extensão, ensino e pesquisa desde 1983, e atua na pesquisa epidemiológica e clínica, principalmente com as classes de animais peçonhentos, agrotóxicos e medicamentos (Cit, 2008). Presta também assistência direta ao paciente, ou auxilia o médico através do telefone nas emergências em todo o estado.

A assistência referida é prestada nas modalidades de atendimento à distância (via telefone) e atendimento in loco, este segundo que pode ser proporcionado a pacientes que se dirigem diretamente até o CIT ou através de encaminhamento através das emergências pediátrico ou adulto.

Este atendimento segue um fluxo padrão – que será descrito detalhadamente mais adiante – onde o plantonista começa então a inquirir o paciente com relação à intoxicação reclamada e toma nota dos detalhes que têm referencia com o ocorrido, efetuando uma espécie de anamnese, entre outros processos necessários anteriores à pesquisa por agentes intoxicantes.

Com a coleta básica de informações cadastrais em mãos, o plantonista começa a fazer a busca pelo conhecimento nas bases de toxicologia para prestar o atendimento correto ao paciente, tendo como fontes de referência um acervo de monografias com informações toxicológicas; um software proprietário do CIT para auxílio ao atendimento e outros sites de referência.

Além destas referências básicas, os plantonistas do CIT muitas vezes se deparam com situações onde o caso do atendimento corrente requerer uma busca mais aprofundada em bases de dados distintas daquelas citadas anteriormente. Estas fontes de pesquisa podem estar em formato digital, contidas em softwares específicos ou na internet, como também podem estar contidas em documentos físicos armazenados nas bibliotecas do CIT.

Neste tipo de ocorrência, o atendimento pode ser prejudicado devido ao fato de que a informação não está devidamente indexada, além da localização geográfica dos itens relacionados se encontrar desfavorável. O plantonista necessita minerar toda esta informação manualmente e estabelecer as relações entre os agentes intoxicantes, para então prestar atendimento ao paciente. Este procedimento pode

acarretar em problemas na administração do conhecimento devido muitas vezes à pressa no momento da busca, e em virtude disso podem ocorrer enganos nos relacionamentos das informações e, posteriormente, no seu armazenamento.

A figura abaixo ilustra brevemente o fluxo de atendimento do CIT (que será apresentado em maior detalhe mais adiante), destacando a tarefa intensiva em conhecimento a ser realizada pelo sistema de conhecimento proposto no andamento deste trabalho.

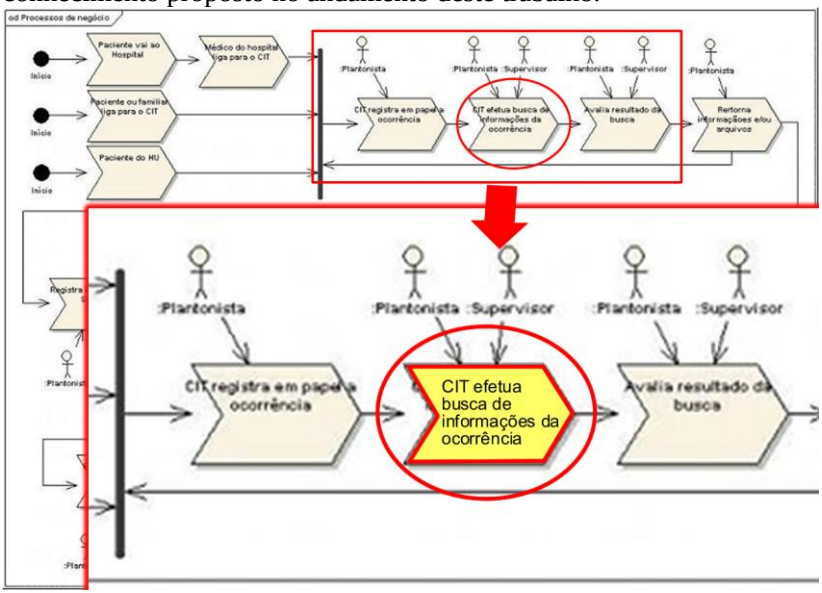


Figura 1 - Fluxo de trabalho de um atendimento no CIT com destaque para a tarefa intensiva em conhecimento a ser trabalhada

1.2. OBJETIVOS

1.2.1. Objetivo Geral

Desenvolver um sistema de recuperação de Conhecimento que permita organizar e recuperar o conhecimento necessário a profissionais de saúde em atendimentos de urgência.

1.2.2. Objetivos Específicos

- Desenvolver uma base de conhecimento sobre agentes tóxicos, medicamentos, animais peçonhentos e tratamentos;
- Desenvolver um mecanismo de busca compatível com as necessidades dos profissionais da área de toxicologia clínica;
- Projetar e Implementar uma interface ergonômica para acesso à base de conhecimento adaptada a dispositivos móveis com interfaces sensíveis ao toque;
- Avaliar a melhor combinação de técnicas de recuperação de informação para composição de estrutura do mecanismo de busca;
- Avaliar qualidade dos resultados e comparar com a prática exercida atualmente no contexto a que o sistema será aplicado.

1.3. JUSTIFICATIVA E MOTIVAÇÃO

O volume de informações utilizadas pelos Centros de Informações Toxicológicas, bem como o número de atendimentos prestados diariamente, é bastante expressivo. É importante ressaltar a existência de um vasto conhecimento disposto heterogeneamente nas dependências físicas e conceituais dos Centros de Informações Toxicológicas em nosso país.

Muitas vezes, existe dificuldade na busca de informações pertinentes a uma determinada substância ativa, pois as fontes de pesquisa estão distribuídas de forma não ordenada em diferentes bancos de dados, bibliotecas, documentos de mídia eletrônica (PDF) ou em sistemas de busca da Internet.

O presente trabalho tem como característica o aprimoramento da manutenção do conhecimento pertinente ao contexto de toxicologia clínica. Deste modo, com a criação de um mecanismo que proporcione a

busca de informações toxicológicas em um banco consolidado de informações pode propiciar atendimento mais veloz devido à indexação deste conteúdo.

Além do maior conforto, os utilizadores podem utilizar este sistema como um auxiliador na tomada de decisões no momento da anamnese e resolução dos atendimentos.

Espera-se que com o desenvolvimento desta pesquisa possibilite-se a homogeneização, armazenamento e posterior recuperação do conhecimento disperso através de sua estruturação e indexação.

Com isso, a recuperação de conhecimento estará centralizada, possibilitando o fácil e rápido acesso além da praticidade de posterior geração de dados estatísticos na possível integração com sistemas desta competência.

1.4. ADERÊNCIA À ENGENHARIA DO CONHECIMENTO

Segundo definição de Alavi e Leidner (2001), sistemas de gestão do Conhecimento são "*sistemas baseados em Tecnologia da Informação desenvolvidos para apoiar os processos organizacionais de criação, armazenamento/recuperação, transferência e aplicação do conhecimento*".

Esta afirmação caracteriza a interdisciplinaridade através do apoio oferecido pelas ferramentas de tecnologia da Engenharia do Conhecimento que é dado a área da Gestão do Conhecimento. Neste trabalho foi desenvolvida uma ferramenta de Engenharia do Conhecimento que visa prover o apoio tecnológico à tarefa intensiva em conhecimento de diagnóstico em na área de Toxicologia Clínica.

A função da ferramenta é gerir o conhecimento aplicado nas atividades realizadas pelos profissionais atendentes dos CIT's, possibilitando o armazenamento, recuperação, transferência e aplicação do conhecimento constante na base gerada a partir deste estudo. Desta forma, é possível afirmar que esta implementação trata-se de uma Ferramenta de Engenharia de Conhecimento para apoio à Gestão do Conhecimento na área da Saúde, desenvolvida sob uma plataforma que permite a disseminação e a centralização do conhecimento em ambiente colaborativo, o qual se encontram os CIT's do Brasil, contexto no qual é aplicado o estudo de deste trabalho.

1.5. ASPECTO METODOLÓGICO

De modo a tornar a contextualização metodológica utilizada durante o processo de desenvolvimento desta pesquisa mais transparente, ilustra-se inicialmente através da fundamentação proposta por Orlikowski & Baroudi (1991), na apresentação dos paradigmas de pesquisa positivista, crítica e interpretativa.

Com auxílio da interpretação de Saunders, Lewis & Thornhill (2002), onde é sugerido que a caracterização da pesquisa está intimamente ligada à pergunta de pesquisa, nota-se após a verificação dos resultados produzidos uma tendência de mescla nos quesitos referentes à distinção qualitativa ou quantitativa da pesquisa, demonstrando um aspecto científico positivista. Segundo Knox, estas são áreas não excludentes mutuamente e portanto, pode-se naturalmente classificar esta pesquisa em mais de um contexto no que diz respeito à caracterização da pesquisa.

Dado este contexto, como primeiro parâmetro para a caracterização temos a seguinte pergunta de pesquisa:
É possível recuperar conhecimento necessário, constante bases de conhecimento dispersas, para atendimento de urgência em saúde?

Torna-se evidente a tendência qualitativa do estudo, considerando-se a pergunta de pesquisa. Todavia, no avanço da revisão de literatura observou-se não só a resposta positiva da pergunta de pesquisa com relação à viabilidade da construção de um motor de busca, como também a possibilidade de utilização de abordagens estruturais distintas, possibilitando a avaliação de comparativo de desempenho entre estas.

A partir desta composição da pesquisa, fez-se necessária a utilização de métricas quantitativas para explorar o grau de diferença de desempenho entre os métodos estudados, de forma a definir de maneira positivista qual o melhor método a ser empregado.

Sob este ponto de vista, é possível definir de maneira geral que o presente estudo trata-se de uma pesquisa **quali-quantitativa** em uma visão **positivista**, partindo-se do problema qualitativo até a avaliação quantitativa efetuada.

Knox (2004) faz uma ligação entre os estágios de aprendizagem propostos por Kolb (1984) ilustrando, a partir destes, a característica indutiva ou dedutiva de um estudo. A tabela abaixo fornece alguns parâmetros para que se possa aclarar a contextualização a partir deste ponto de vista:

Tabela 1 - Ligação entre ciclo de aprendizagem de Kolb's e Interpretação Indutiva e Dedutiva
Fonte: Adaptada de (Knox, 2004)

Ciclos de aprendizagem de Kolb	Indução / Dedução
Experiência Concreta (Sentir) – o aprendizado ocorre pela imersão no problema, onde há mais confiança na intuição que na lógica	Indução
Observação Reflexiva (Observar) – considera os experimentos prévios, refletindo para formular expectativas	Indução / Dedução
Conceitualização Abstrata (Pensar) – análise do problema, reflexão para formular novas teorias para o futuro	Indução / Dedução
Experimentação Ativa (Fazer) – aplicação de pensamentos e idéias, aprendendo através de tentativa e erro	Dedução

Partindo-se do princípio que o problema a ser resolvido neste estudo é clássico em Engenharia do Conhecimento, considerava-se a hipótese de trabalhos relacionados terem utilidade na reutilização de aplicações desenvolvidas para fins semelhantes. Com isso, os experimentos realizados previamente à realização dos primeiros experimentos possibilitaram a formação de expectativas com relação aos resultados, caracterizando o caráter **Indutivo/Dedutivo** da pesquisa, conforme a tabela previamente apresentada.

Com relação ao fluxo adotado na realização do trabalho, os procedimentos determinados para a realização foram determinados, na ordem:

- Levantamento das características do problema por meio de entrevistas com os profissionais atuantes na área de estudo de caso;
- Levantamento das aplicações específicas utilizadas para resolução deste tipo de problema, não contextualizadas apenas na área de enfoque;
- Levantamento bibliográfico de estudos científicos relacionados ao enfoque utilizado para resolução do problema de heterogeneidade, dispersão e recuperação do conhecimento;
- Contextualização dos enfoques não relativos à conjuntura do estudo de caso;

- Implementação e avaliação parcial dos primeiros experimentos com integração das bases de conhecimento e recuperação do conhecimento na base resultante;
- Avaliação dos diferentes métodos utilizados para resolução do problema.

A realização de cada uma das etapas teve como parâmetro as delimitações previamente acertadas e descritas na seção posterior, tendo em vista a não dispersão de foco e viabilização da completude do estudo.

1.6. ESCOPO E DELIMITAÇÃO DO TRABALHO

- A pesquisa abrange o uso de ontologias como instrumentos para representação do domínio de conhecimento;
- A partir deste estudo é possibilitada a integração de bases de dados heterogêneas seguindo uma classificação previamente especificada por profissionais da área de toxicologia;
- Este trabalho implementa uma estrutura modular integrada para possibilitar a pesquisa de agentes tóxicos de diversas fontes em uma única interface;
- Os resultados encontrados são contabilizados e comparados a partir de métricas quantitativas específicas para as finalidades apresentadas, além de um comparativo quantitativo com outras abordagens relacionadas;
- Não serão pesquisadas ou utilizadas outras técnicas como sistemas de inteligência artificial ou sistemas de informação;
- O contexto do conhecimento envolvido é o de Toxicologia Clínica, incluindo apenas agentes tóxicos, animais peçonhentos e tratamento de intoxicações;
- As fontes de conhecimento utilizadas no desenvolvimento da base de conhecimento incluem o UMLS Metathesaurus, HSDB, DeCS, bases da Anvisa e repositório do CIT/SC.

1.7. ESTRUTURA DO TRABALHO

A definição da estrutura deste trabalho visa abordar o aspecto teórico no intuito de contextualizar o problema de caráter genérico na área de Engenharia do Conhecimento, também visualizado na área da saúde. Na seção introdutória deste documento estão caracterizados brevemente os

fatores mais importantes relacionados ao contexto e definição do problema, sendo aprofundadas na seção de Fundamentação Teórica.

A fim de garantir a consistência de conteúdo e metodologia adequadas à pesquisa, realizou-se uma REVISÃO SISTEMÁTICA DA LITERATURA, constante no capítulo 2, abordando os contextos de TI em Toxicologia e a Integração de Bases de Dados Heterogêneas.

No capítulo 3 é apresentado o contexto do estudo de caso realizado, na subárea de Toxicologia Clínica, através da análise do ambiente dos CIT's. Nesta mesma seção são apresentados também conceitos sobre Engenharia do Conhecimento, como o seu surgimento, evolução, diferentes abordagens, a importância do conhecimento como fator de valor agregado, Sistemas de Conhecimento e ferramentas e métodos para sua construção.

Também está contida neste capítulo uma apresentação de trabalhos relacionados utilizados para a resolução de problemas semelhantes ao contexto deste trabalho, abordando áreas de web semântica, busca por linguagem natural, busca e indexação por arquivos invertidos e expansão de consulta.

A indicação de resolução do problema é apresentada no capítulo 4 - PROPOSTA, e contém descrições detalhadas partindo da conceitualização modular do protótipo, demonstrando as funcionalidades básicas exercidas por cada módulo, traçando um paralelo à utilização conceitual de algumas figuras apresentadas na metodologia CommonKADS, seguida da descrição do processo de desenvolvimento da Base de Conhecimento, módulos de conversão (*Knowledge Convertor*), Módulo de Aprimoramento Semântico, Expansão de consulta e Interface.

Ao final do capítulo 5 são apresentadas algumas considerações feitas sobre o desenvolvimento e funcionalidades do protótipo aplicado, traçando um relacionamento introdutório com os resultados e as diversas abordagens estruturais avaliadas durante o estudo. A Figura 2Figura 2 aponta a inter-relação entre os capítulos do presente trabalho ilustrada através de um mapa conceitual.

No capitulo 6 - RESULTADOS, é apresentada primeiramente uma breve conceitualização do método de avaliação GQM (*Goal, Question, Metric*), seguindo da ilustração da definição dos objetivos, questões e métricas utilizadas para a análise dos resultados obtidos com os experimentos do protótipo desenvolvido. Em seguida é demonstrado o contexto onde foram realizados os experimentos, bem como a caracterização dos elementos utilizados no protótipo, entre outros itens relacionados.

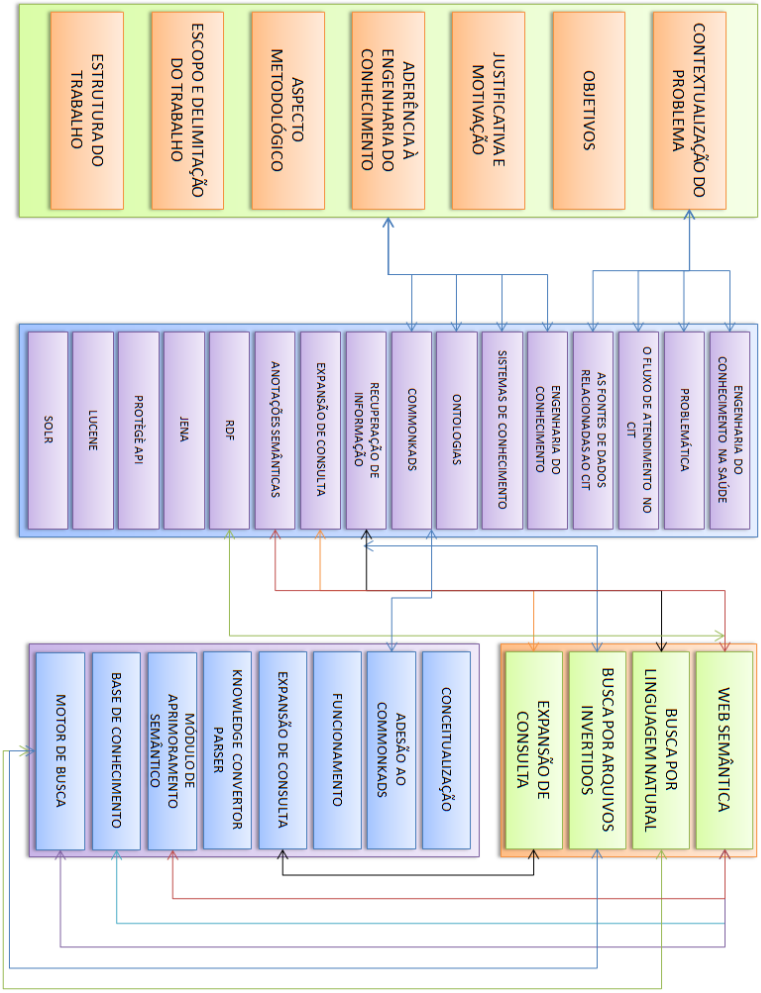


Figura 2 – Mapa conceitual da estrutura dos pontos do trabalho

Subsequentemente são interpretados e analisados os resultados obtidos através da realização dos experimentos, seguidos da seção de discussão, que contém a Avaliação dos Resultados e Implicações, Ameaças à Validação e Inferências com o desenvolvimento do trabalho. E finalmente, são apresentadas as conclusões do trabalho, contendo sugestões para trabalhos futuros e itens relacionados.

2. REVISÃO SISTEMÁTICA DA LITERATURA

Neste capítulo serão ilustrados assuntos e trabalhos desenvolvidos que dizem respeito ao objeto de pesquisa deste estudo. Para isso foi realizada uma revisão sistemática da literatura visando contemplar nesta pesquisa os estudos realizados no meio acadêmico relativos ao desenvolvimento de **aplicações de Tecnologia da Informação para apoio às atividades realizadas por profissionais da área da Saúde**, mais especificamente na área de **Toxicologia Clínica**. A revisão realizada contempla também trabalhos relacionados ao desenvolvimento de **aplicações para integração de bases de dados/conhecimento heterogêneas**.

As especificações de termos e mecanismos de busca utilizados para realização destas pesquisas, bem como a descrição dos trabalhos explorados são expostos a seguir:

2.1. APLICAÇÕES DE TI EM APOIO À SAÚDE

A utilização de aplicações de Tecnologia da Informação em prol do melhoramento do desempenho na realização das atividades traz consigo marcos históricos da Tecnologia da Informação, visto que foram os primeiros experimentos realizados na área de Inteligência Artificial.

Aplicativos como MYCIN (Shortliffe, 1976) são responsáveis por este registro histórico, que será mais bem detalhado entre outros aspectos na seção Tipo de comunicação.

Com relação à pesquisa efetuada para mineração das iniciativas de TI em Saúde, utilizou-se um método de revisão sistemática da literatura proposto por (Kitchenham, 2004). Especificamente para a área de Aplicações de TI em Saúde, foram elencados os indexadores PUBMed, ScienceDirect e BioMED Central.

Os termos selecionados para este contexto tinham predominantemente relação com “Recuperação de Informação”, “Sistemas de Informação Clínica” e “Centro de Informações Toxicológicas" (em inglês, “poison centers”).

A partir destes termos, foram encontrados inicialmente 393 artigos nos mecanismos indexadores citados. Após a exploração destes artigos, foram selecionados 33 artigos para exploração aprofundada, devido à identificação de similaridade com este trabalho.

Os parâmetros utilizados para pesquisa em cada um dos indexadores podem ser vistos na Tabela 2.

Tabela 2 - Parâmetros utilizados para pesquisa de IT em toxicologia

Indexador	Parâmetros	Artigos retornados
PUBMed	("information retrieval" OR "clinical information systems") AND ("toxicology" OR "poison center" OR "poison centre")	113
BioMed Central	("information retrieval" OR "clinical information systems") AND ("toxicology" OR "poison center" OR "poison control center" OR "poison centre")	15
Science Direct	pub-date > 1979 and ALL(("information retrieval" OR "clinical information systems") AND ("toxicology" OR "poison center" OR "poison centre"))	265
Total	-	393

A lista completa de todas as classificações dos trabalhos descritos nesta seção, bem como uma breve descrição de cada trabalho recuperado está disponibilizada no apêndice tal.

2.1.1. Tipos de abordagem

Para organização dos trabalhos selecionados, identificou-se uma categorização específica para classificação dos trabalhos quanto à sua abordagem, visto que alguns autores contemplavam Sistemas de Apoio à Diagnóstico em sua completude, enquanto outros se ativeram mais profundamente às questões inerentes à bases de conhecimento. As categorizações identificadas durante este estudo podem ser visualizadas na **Tabela 3**.

Tabela 3- Categorizações das abordagens identificadas no estudo

Categoria	Sigla
Clinical Information System /Sistema de apoio a diagnostico/Sistemas Especialistas	**(**CIS/CDSS/SE**)**
Base de conhecimento	(BC)
Mashup de bases de conhecimento	(MDB)
Algoritmos para busca em bases de conhecimento	(ABBC)
Técnicas de mapeamento do conhecimento	(TMCP)

contido em publicações

A distribuição destes estudos nas abordagens categorizadas acima pode ser identificada nos gráficos por popularidade e número de trabalhos, ilustrados em Figura 3 e Figura 4.

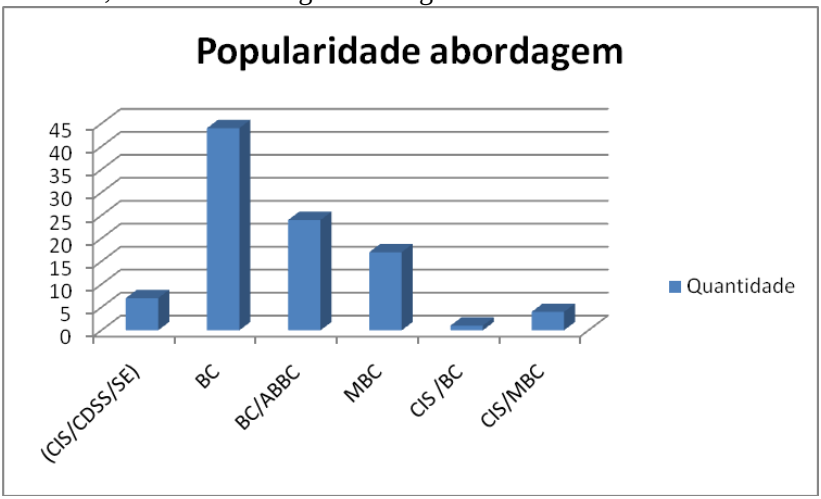


Figura 3 - Distribuição de abordagens por popularidade

Figura 4 - Distribuição de abordagens por número de trabalhos

A primeira característica que pode ser visualizada e **deve ser observada** nestes dois gráficos é que eles não apresentam a mesma proporção de trabalhos, diferenciando o quesito "popularidade" e

"número de trabalhos". Isto acontece devido à forma com que foram organizados os dados recolhidos durante a pesquisa: considera-se um fator que **incrementa** a **popularidade** o número de trabalhos científicos encontrados que se utilizam de uma determinada abordagem para efetuar estudos, comparações, avaliações em que haja uma consideração desta abordagem. Por outro lado, para o **incremento** do **número de trabalhos** de uma determinada abordagem, este sim necessita estar trabalhando diretamente com o tema, seja utilizando como apoio ao desenvolvimento de um trabalho, ou então que esta abordagem esteja sendo descrita como desenvolvimento na publicação científica.

Observados estes detalhamentos e novamente visualizando os gráficos, pode-se identificar também que são amplamente discutidos no meio acadêmico trabalhos que visam a composição/estruturação de bases de dados/conhecimento representados no gráfico pela sigla **BC**, responsáveis nesta revisão sistemática por mais de 1/3 das publicações. Entre estas abordagens, podemos citar:

- CCRIS (Wexler, 2001; Nlm, 2003);
- ChemIdPlus (Berman *et al.*, 1992; Wexler, 2001; Nlm, 2003);
- GENE-TOX (Wassom, 1985; Wexler, 2001; Nlm, 2003);
- HSDB - Hazardous Substances Data Bank (Fonger *et al.*, 2000; Wexler, 2001; Nlm, 2003)
- TRI - Toxics Release Inventory (Wexler, 2001; Nlm, 2003)
- CTD - Comparative Toxicogenomics Database (Mattingly *et al.*, 2003);
- DRUGDEX, EMERGINDEX, IDENTIDEX, POSINDEX (Lundsgaarde e Moreshead, 1991; Micromedex, 2010);
- IRPTC (Huismans, 1980; Kurlyandskiy e Sidorov, 2003);
- MEDLARS (Sodergren, 1973; Lindberg *et al.*, 1993; Robinson *et al.*, 2000; Wright, 2001);
- RTECS - Registry of Toxic effects of Chemical Substances (Rtecs, 2000; Kurlyandskiy e Sidorov, 2003) e
- TRACE (Anderson *et al.*, 2000; Robinson *et al.*, 2000).

Este tipo de trabalho está associado a estudos que visam a geração e manutenção de conhecimento em para disponibilização integrada para a utilização em pesquisas na área da saúde. Muitas destas bases de conhecimento já estão integradas em algo que se pode chamar de **meta-Tesauro**, no sentido de que esta passa a estar em um nível hierárquico superior, possibilitando o acesso à diversas áreas do conhecimento através de mecanismos integrados disponibilizados para este fim.

A possibilidade de utilização desta abordagem de maneira integrada, a qual neste trabalho foi chamado *mashup*, é representada por sua popularidade (vide Figura 3) em meio aos trabalhos acadêmicos aqui abordados.

Dentre os trabalhos relacionados a esta abordagem reunidos nesta revisão sistemática, podemos citar UMLS (Bodenreider e Burgun, 2004), TOXNET (Wexler, 2001; Nlm, 2003), TOXLINE (Wexler, 2001; Nlm, 2003) e MICROMEDEX (Micromedex, 2010). A estrutura hierárquica destes meta-Tesauros e suas bases de conhecimento caracterizam-se como segue:

- UMLS
    - TOXNET
        - HSDB
        - CCRIS
        - GENE-TOX
        - TOXLINE
            - EMIC
            - DART / ETIC
            - TRI
            - ChemIdPlus
- MICROMEDEX
    - POSINDEX
    - DRUGDEX
    - EMERGINDEX
    - IDENTIDEX

2.1.2. Áreas de interesse

Também durante o processo de exploração da documentação encontrada, observou-se a caracterização de uma categorização por área de interesse destas bases de conhecimento. No gráfico da figura, podemos visualizar quais são as áreas mais publicadas encontradas nos documentos recuperados nesta revisão.

Assim como no caso da categorização por abordagem, é possível observar a diferença entre o número de trabalhos publicados em relação ao índice de popularidade obtido a partir do número de citações de um determinado trabalho. Seguindo a mesma regra já adotada na análise sobre as abordagens, verifica-se que os trabalhos encontrados têm mais referências às áreas de produtos químicos / agentes tóxicos e medicina. Observa-se também uma deficiência em publicações que façam referência a produtos agrotóxicos, controle ambiental e segurança

da saúde, no que tange bases de conhecimento a serem utilizadas pelos Centros de Informações Toxicológicas.

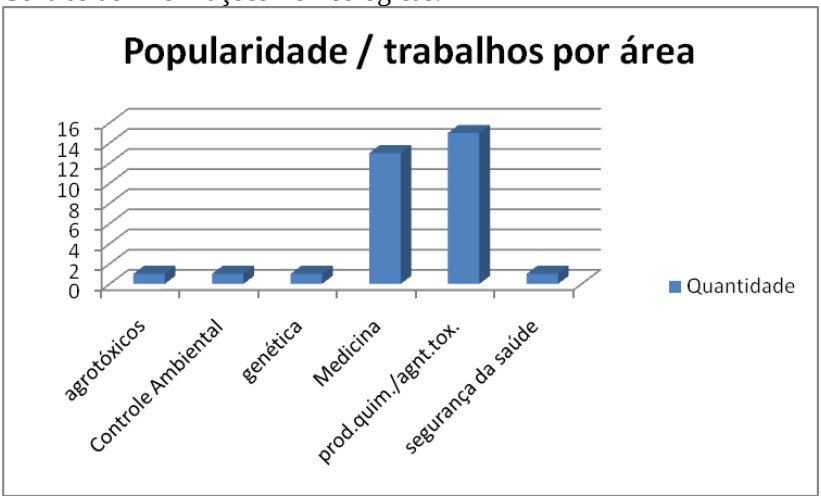


Figura 5 - Popularidade de trabalhos publicados por área

Figura 6 - Quantidade de trabalhos publicados por área

As categorizações que são utilizadas nos gráficos das figuras 4 e 5 são fruto da identificação de cada um destes aspectos contidos nos trabalhos obtidos na revisão sistemática.

A seguir são descritos os trabalhos constantes em cada uma das categorias encontradas, cuja lista completa acompanhada das referências bibliográficas está disponível no apêndice deste documento.

Tabela 4 - Trabalhos por área

- Agrotóxicos
  - EMIC
- Controle ambiental
  - TRI - Toxics Release Inventory
- Genética
  - GENE-TOX
- Medicina
  - botXminer
  - CoPub Mapper
  - MCA
  - Mycin
  - INTERNIST-I
  - CCRIS
  - EMERGINDEX
  - MEDLARS
  - MeSH
  - MEDLINE
  - EMBASE
- Produtos químicos / agentes tóxicos
  - HT-Attending
  - DBX
  - HSDB
  - DART / ETIC
  - ChemIdPlus
  - CTD
  - POSINDEX
  - DRUGDEX
  - IDENTIDEX
  - TRACE
  - IRPTC
  - RTECS
  - TOXBASE
  - TOXNET
  - TOXLINE
- Segurança da saúde
  - IPSP thesaurus.

2.1.3. Sistemas de Conhecimento em Saúde

Foram encontrados também neste estudo ocorrências de implementações de sistemas de conhecimento aplicados à área da saúde. Alguns destes fazem parte do contexto histórico, identificados nos primórdios da inteligência artificial no intuito de prover apoio em tarefas de diagnóstico, em que se pode citar, por exemplo, o MYCIN (Shortliffe, 1976).

Muitas das iniciativas das instituições que vislumbravam a composição de bases de conhecimento também desenvolveram métodos que possibilitassem a pesquisa nestas bases, criando sistemas de conhecimento que pudessem minerar o conteúdo produzido. Muitas destas bases e *mashups* de Conhecimento - em que se pode fazer referência às iniciativas da Biblioteca Nacional de Medicina dos Estados Unidos (Nlm, 2003), assim também como iniciativas privadas como Micromedex - disponibilizam este tipo de mecanismos.

Existem pesquisas também na área de desenvolvimento de algoritmos de mineração em bases de conhecimento de conteúdo científico, que visam otimizar a forma de pesquisa nestes locais. Exemplos destes algoritmos são botXminer (Mudunuri *et al.*, 2006) e CoPub Mapper (Alako *et al.*, 2005), que estão entre os algoritmos desenvolvidos para mineração de conhecimento na base de artigos científicos sobre medicina MEDLINE (Sewell e Bevan, 1976; Kenton e Scott, 1978; Lindberg *et al.*, 1993; Ludl *et al.*, 1996; Gehanno *et al.*, 1998; Anderson *et al.*, 2000; Robinson *et al.*, 2000; Suarez-Almazor *et al.*, 2000; Cimino *et al.*, 2003; Alako *et al.*, 2005; Alpi, 2005; Darmoni *et al.*, 2006; Mudunuri *et al.*, 2006).

O desenvolvimento de aplicações para integração de bases de conhecimento é uma atividade utilizada para fins genéricos. Na seção seguinte serão abordados estudos que tratam destes fins, utilizando uma visão de mundo holística a fim de obter conhecimento sobre alternativas para estruturação deste tipo de trabalho.

2.2. APLICAÇÕES PARA INTEGRAÇÃO DE BASES DE DADOS HETEROGÊNEAS

Este tipo de aplicação tem como objetivo encontrar maneiras para possibilitar ao usuário pesquisar em um universo de conhecimento disperso e heterogêneo o conteúdo desejado de maneira transparente e centralizada.

O método utilizado para mineração dos estudos relacionados à integração de bases de conhecimento foi semelhante ao anterior, de Barbara Kitchenham. Especificamente para esta área, os mecanismos indexadores utilizados para pesquisa foram: ScienceDirect, IEEExplore e ACM Digital Library.

Os termos selecionados para este contexto tinham predominantemente relação com "integração de bancos de dados heterogêneos" e "integração de dados/base de conhecimento/ontologias", filtrando por artigos publicados a partir do ano de 2005.

Partindo destes termos, foram encontrados inicialmente 1171 artigos nos mecanismos indexadores citados. Após a exploração destes artigos, foram selecionados 28 artigos para exploração aprofundada, devido à identificação de similaridade com este trabalho.

Tabela 5 – Parâmetros utilizados para pesquisa em integração de bases de dados heterogêneas

Indexador	Parâmetros	Artigos retornados
ACM (Journals)	(heterogeneous and database and integration, and Data and integration, and knowledge and base, and ontology, and heterogeneous and databases) and (PublishedAs:journal) and (FtFlag:yes)	73
IEEExplore	Publication Year: 2005 - 2011 heterogeneous knowledge base integration Publication Year: 2005 - 2011	163
IEEExplore (Journals)	Content Type: Journals Publication Year: 2005 – 2011 (((“heterogeneous database integration”) OR (“Data integration, knowledge base”) OR ("ontology") AND (“heterogeneous databases”)))	15

Science Direct	pub-date > 1999 AND (“heterogeneous database integration”) AND (“Data integration, knowledge base, ontology”) AND (“Data integration”) AND (“knowledge base”) AND (“heterogeneous databases”)	920

Os parâmetros utilizados para pesquisa em cada um dos indexadores podem ser visualizados na **Tabela 5**.

A lista completa de todas as classificações dos trabalhos descritos nesta seção, bem como uma breve descrição de cada trabalho recuperado está disponibilizada no apêndice tal.

2.2.1. Atributos de categorização dos trabalhos

Para esta revisão sistemática da literatura foram adotados alguns parâmetros de observação através de atributos identificados para cada trabalho. Estes atributos dizem respeito aos requisitos levantados para uma possível estruturação de um mecanismo para resolução do problema da integração de bases de dados/conhecimento heterogêneas.

A Tabela 6 classifica e explana os atributos avaliados, para posterior apreciação detalhada.

Tabela 6 – Atributos observados nos trabalhos pesquisados

Parâmetro	Descrição
Tipo de integração	Trata da maneira com que é feita a integração, se existe ou não uma extração para armazenamento do conhecimento em uma base local.
Tipo de comunicação	Trata da maneira oferecida pela estrutura para comunicação com entidades de software externas.
Área de aplicação	Trata da área de aplicação de um determinado estudo.
Tipo de avaliação de resultados	Trata da maneira com que foram avaliados os resultados dos experimentos realizados sobre um determinado estudo.
Ampliação do escopo de pesquisa	Trata da implementação de mecanismos que colaboram com a ampliação do escopo da pesquisa.

A partir destes critérios de observação Foi feita a extração dos dados referentes aos trabalhos coletados durante a pesquisa dos artigos científicos, conforme será ilustrado nos seguintes tópicos:

2.2.2. Abordagem por tipo de integração

Durante a avaliação do conteúdo das publicações acessadas para o contexto da pesquisa sobre mecanismos de integração de bases de dados heterogêneas, foram identificadas duas grandes classes para abordar a metodologia da referida integração, das quais podemos cunhar como **integração distribuída** ou **bases de dados replicadas localmente**.

Cada uma destas abordagens possui uma aplicação referente ao contexto de utilização. Na maioria dos casos, mesmo que possam ser utilizadas ambas as soluções, existem fatores chave que determinam qual a melhor opção de integração.

Na sequência, é mostrado um gráfico que reflete o uso de cada uma das abordagens no contexto dos artigos retornados para esta revisão de literatura.

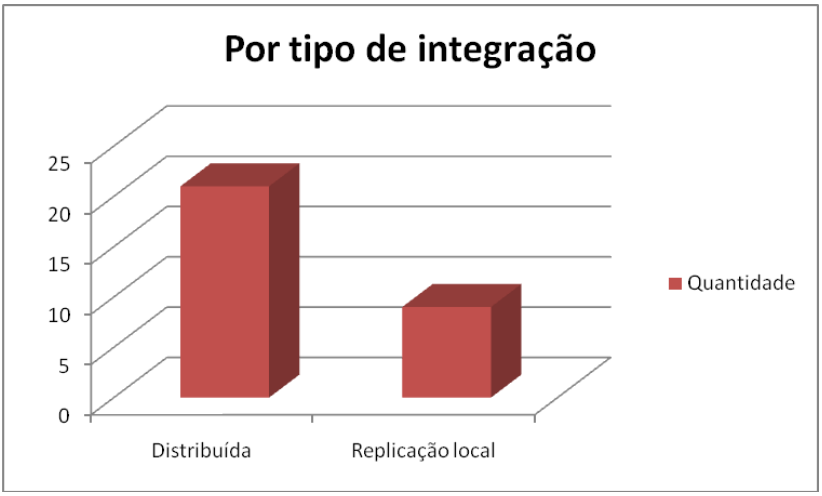


Figura 7 - Proporção de trabalhos por abordagem de integração

É possível visualizar que a maioria dos trabalhos recuperados trata da integração de bases de dados/conhecimento heterogêneas de maneira distribuída. Isso se justifica pela aplicabilidade desta abordagem no contexto do conhecimento que é atualizado constantemente, onde se torna prioridade a recuperação do conhecimento atualizado em detrimento da performance da busca.

Ao contrário, a replicação local do conhecimento constante em bases de dados heterogêneas apresenta deficiências em sua atualização, porém colabora para o aperfeiçoamento da performance no tempo das buscas, considerando também a maior disponibilidade no caso de um sistema que possa ter eventual interrupção da comunicação causada por

problemas na rede.

2.2.3. Tipo de comunicação

Também durante a exploração do conteúdo dos artigos científicos recuperados nesta revisão, foram encontrados trabalhos que possuíam diversos modelos de comunicação com a sua interface de manipulação. Dentre estes trabalhos, foram identificadas as seguintes categorizações básicas: **independente** (utilizando SPARQL), **independente** (utilizando recursos do framework JADE) e formato proprietário, além dos trabalhos que não identificaram o mecanismo utilizado para comunicação com a interface.

Baseado nesta categorização de trabalhos é possível identificar a proporcionalidade da utilização destas abordagens no gráfico da Figura 8.

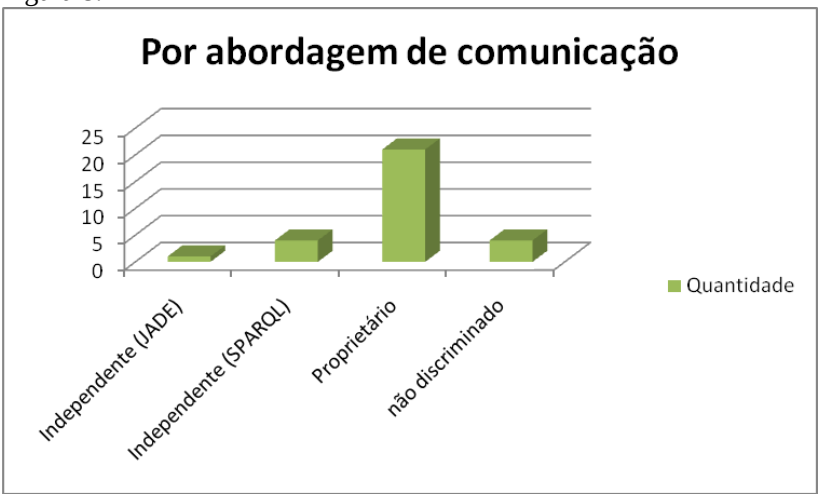


Figura 8 - Proporção de trabalhos por tipo de abordagem de comunicação com a interface de manipulação

Através desta visualização deste gráfico sobre abordagens de comunicação e, levando em consideração os trabalhos selecionados para observação, é possível concluir que na maioria dos estudos científicos relacionados ao contexto de integração de bases de dados heterogêneas não existe uma preocupação com a reutilização destes sistemas por diversos formatos de interface de manipulação.

É possível inferir que isto aconteça pelos autores não terem intenção da aplicabilidade de seus estudos para múltiplos ambientes, ou

por não haver interesse da disponibilização deste tipo de interface para fins não comerciais.

2.2.4. Áreas de aplicação

Na seção **Aplicações de TI em apoio à saúde**, o objetivo foi encontrar trabalhos relacionados à aplicação de tecnologia da informação na área da saúde, mais especificamente, toxicologia clínica. Dentre estes trabalhos, abordagens sobre integração de bases de dados heterogêneas neste contexto também foram encontradas.

De modo a compreender a ilustração de trabalhos sobre este assunto abordando outras áreas de aplicação, foram encontrados nesta segunda seção da revisão sistemática da literatura estudos aplicados em diversas áreas, entre elas: agricultura, e-business, data warehouse, documentação de projetos, informações biomédicas, aplicações ponto-a-ponto e aprendizado virtual.

A distribuição dos trabalhos científicos para cada área identificada neste estudo é ilustrada no gráfico da Figura 9.

Figura 9 - Proporção de trabalhos por área de aplicação

É possível facilmente visualizar que os estudos científicos realizados com o intuito da integração de bases de dados heterogêneas têm abordado em sua maioria a resolução deste problema no que diz respeito a informações biomédicas, assim como aplicações de conhecimento sobre aplicações de conhecimento genérico.

Isso caracteriza a importância de se trabalhar em contexto genérico, ilustrando claramente a aplicação dos conceitos básicos de

engenharia do conhecimento nestes estudos. A área de informações biomédicas tem referência direta com o objeto de pesquisa do presente trabalho, tendo relação íntima com as características funcionais a serem desempenhadas na proposta.

2.2.5. Tipo de avaliação dos resultados

Observar a metodologia de avaliação de resultados dos estudos utilizados como parâmetro na pesquisa torna-se ponto importante no sentido de verificar a profundidade da carga da relação entre teoria e prática, a fim de caracterizar o grau de prioridade de cada abordagem a ser considerada uma opção para o desenvolvimento de um trabalho.

Neste sentido, foram identificados graus de classificação para os tipos de avaliação encontrados nos estudos recuperados através desta revisão. Estes graus são:

- Análise comparativa dos resultados;
- Discussão teórica;
- Ilustração dos artefatos produzidos;
- Estatísticas quantitativas e
- outros métodos quantitativos.

Figura 10 - Proporção de trabalhos por tipo de resultado

2.2.6. Ampliação do escopo da pesquisa

Finalmente, um dos requisitos levantados para a implementação da

proposta para este trabalho é a possibilidade da expansão do escopo de uma determinada pesquisa.

Visando identificar as diversas alternativas para a utilização de ampliação deste escopo, foram identificados trabalhos que caracterizam as abordagens que podem ser visualizadas na figura.

As abordagens encontradas para expansão da consulta foram:

- Clicklogs – trata-se de uma abordagem que identifica o perfil do usuário levando em consideração o seu comportamento em relação aos links que são utilizados para chegar à uma determinada informação (Nandi e Bernstein, 2009);
- Módulo semântico – qualquer abordagem que utilize técnicas de aperfeiçoamento semântico para ampliação da abordagem de consulta;
- Navegação em árvore – técnicas de disponibilização da árvore de classificação do conhecimento para navegação em interface;
- Técnicas de IA – técnicas de raciocínio automatizado através de *reasoners* para expansão do escopo de pesquisa;
- Expansão/tradução de consulta – utilização de bases de dados/conhecimento auxiliares para composição/concatenação de termos e consultas para ampliação do contexto da consulta;
- Não apresentado – trabalhos que não apresentaram nenhum mecanismo de expansão de consulta.

Figura 11 - Tipos de abordagem para expansão de escopo de pesquisa

3. FUNDAMENTAÇÃO TEÓRICA

3.1. ENGENHARIA DO CONHECIMENTO NA SAÚDE

Segundo Heloise Manica (Manica *et al.*, 2009), o "*conhecimento está inserido na maioria das tarefas executadas por profissionais em saúde e raramente surge de forma isolada da atividade*". Esta autora afirma ainda que estas atividades podem ser divididas em tarefas, que podem também ser classificadas como "tarefas intensivas em conhecimento", que representam passo-a-passo uma atividade normalmente realizada por seres humanos.

As atividades relacionadas a diagnóstico, realizadas frequentemente na área da saúde, são tarefas tipicamente mapeáveis através da Engenharia do Conhecimento, estando esclarecidas em muitas das metodologias desta disciplina.

Neste contexto, o relacionamento entre Engenharia do Conhecimento e a área da saúde é justificada conforme a afirmativa de Landry (2006), quando ele comenta sobre esta cooperação:

> "Knowledge management studies tend to adopt the organization as their focus of attention, thus looking at how organizational characteristics affect the translation and implementation of knowledge in the solving of public health problems. […] For public health organizations […], the capability to acquire, create, share and apply knowledge represents their most significant capability in terms of solving public health problems" (Landry *et al.*, 2006).

Desta forma, Landry caracteriza a forma de colaboração da Engenharia do Conhecimento na área da saúde citando as atividades que a EC objetiva atuar de maneira genérica. Um dos tipos de Sistemas de Conhecimento representados na área da saúde são os *Clinical Decision Support Systems* – CDSS (Sistemas de Suporte à Decisão Clínica).

Eis a seguir algumas das definições apresentadas para CDSS:

- Musen define um CDSS como sendo qualquer peça de software que obtém informação sobre uma situação clínica como entrada e produz inferências sobre as saídas que podem auxiliar profissionais na tomada de decisões (Musen e Van Bemmel, 1997);

- Miller e Geissbuhler definem CDSS como sendo um provedor de suporte à diagnóstico como um algoritmo baseado em computação que auxilia o profissional com um ou mais componentes no processo de diagnóstico (Miller e Geissbuhler, 1999);
- Sim (2001) define CDSS como sendo "softwares destinados a auxiliar na tomada de decisão clínica, na qual as características de cada paciente são compatíveis com uma base de conhecimento clínica do paciente e avaliações específicas ou recomendações são apresentadas para o médico ou o paciente para uma decisão.

Este tipo de sistema é particularmente o foco em que o atual estudo é inserido, apresentando algumas peculiaridades, mas mantendo-se no contexto da área da saúde e suporte à decisão. A seguir, será descrita a utilização de Engenharia do Conhecimento na resolução de problemas de dispersão e heterogeneidade do Conhecimento, cuja proposta de solução será descrita mais adiante.

3.2. DESAFIO

A heterogeneidade e dispersão de bases de dados e conhecimento são fatores que atingem diversos domínios, entre as quais podemos citar: Direito, Administração, Medicina, entre outras. A "*falta de padronização na representação deste conhecimento dificulta a compreensão por parte de terceiros*" (Farias *et al.*; Chang *et al.*, 1998; Scotney e Mcclean, 1999; Tsoumakas *et al.*, 2004).

Euzenat (2007) ilustra a ocorrência deste problema levando em consideração o tipo de conhecimento utilizado no contexto de trabalho com livros:

> "[…] two organisations dealing with books: one is a cultural product electronic commerce site (which sells books, music, movies, etc.) and the other is a university library. The activities of both organisations deal with some related products, the books, but are concerned with different aspects of these: the seller is concerned by the margin, the publisher or the type of binding. The library, in turn, pays more attention to the topic, the size and the year of publication. Both are concerned by the price and the author. Yet they may consider these

> differently, because the price can include tax and shipping fees or not and being expressed in different currencies or because the authors can be denoted by individual objects or by the character string of their names. Moreover, the seller may organise the books according to their commercial types and the library according to their literary types. In summary, these two organisations will obviously have different and heterogeneous ontologies […]" (Euzenat *et al.*, 2007).

No início dos anos 90, Worboys e Deen fizeram menção sobre a distribuição geográfica de bancos de dados e a complexidade das estruturas de dados contidas em cada localização, comentando sobre a dificuldade de uma possível integração devido à problemas de heterogeneidades semântica ou física (Worboys e Deen, 1991; Chang *et al.*, 1998), enquanto que Chung (1990) aborda o impacto da falta de compartilhamento de dados nas organizações:

> "[…] there can be several different DBMSs in a data center. Currently, there are no effective means to share these heterogeneous databases. The lack of effective data sharing causes inefficient engineering and manufacturing activities and business operations. Duplicated data at different locations results in data inconsistency. The development of the same applications in different data manipulation languages used by different DBMSs incurs unnecessary human cost" (Chung, 1990).

É possível afirmar que, nos dias de hoje, a maioria das grandes organizações têm um grande número de fontes de conhecimento, distribuídas em nodos de suas redes de sistemas de informação. No momento em que estas organizações têm a intenção de integrar o conhecimento destas fontes distribuídas, são necessárias técnicas específicas que alguns autores chamam de "fusão de conhecimento" (Chang *et al.*, 1998; Scotney e Mcclean, 1999; Preece *et al.*, 2000).

Preece explica ainda que uma consulta simples em um banco de dados distribuído recupera as instâncias de dados, porém, isto não significa que está acontecendo a fusão do conhecimento, pois estas não estão sendo associadas baseadas no contexto, sem a informação de como estas devem ser interpretadas ou utilizadas.

> "A distributed database query retrieves data instances, but this is not enough for knowledge fusion. To combine information in a meaningful

> way, data instances need associated knowledge of their context: how they should be interpreted and how they can be used" (Preece *et al.*, 2000).

Este problema também é encontrado na área da saúde. Conforme afirmativa de (Chu *et al.*, 1995), a pesquisa e a prática da medicina têm requerido implementações avançadas em recursos de gerenciamento de bancos de dados.

Afirma-se que isto é causado pela redução dos custos computacionais e a abrangência do acesso à internet, que cria um oceano de dados eletrônicos, gerados através natureza descentralizada da comunidade científica, o que resulta em uma miscelânea heterogênea de implementações de bancos de dados. Isto sugere que também nesta área o acesso e agregação entre as bases de dados e conhecimento seja dificultosa (Sujansky, 2001).

Chu considera primeiramente a necessidade do acesso a dados de pacientes através de sistemas de informação, e cita os HIS – *hospital information systems*, RIS – *radiology information systems*, PACS – *Picture archiving and communication systems*, entre outros.

Neste contexto se insere o cotidiano dos Centros de Informações Toxicológicas. Esta é uma subárea da saúde recebe por herança o mesmo problema causado pelo excesso de conhecimento relacionado e não integrado (Lovell e Celler, 1999; Landry *et al.*, 2006; Cabral *et al.*, 2008; Cabral *et al.*, 2009; Ribeiro *et al.*, 2009). Os profissionais que se utilizam do conhecimento prévio gerado através de estudos empíricos, locais ou não, não têm um mecanismo eficiente no qual possam efetuar as pesquisas de maneira rápida e prática, o que é primordial no atendimento de urgência.

As fontes de conhecimento necessárias para a resolução de problemas relacionados à Toxicologia Clínica estão dispersas em fontes de conhecimento físicas ou mecanismos disponíveis na internet, cada um com sua peculiaridade, o que faz com que o profissional atendente tenha que, muitas vezes, efetuar pesquisas alternando entre estes mecanismos, a fim de encontrar o item que melhor se adéqua às necessidades (Ribeiro *et al.*, 2009).

Traçando um paralelo com o segundo ponto abordado por Chu, que diz respeito à "necessidade de pesquisas que não sejam tradicionais no que diz respeito aos parâmetros de entrada", em que exemplifica mencionando os atributos "paciente, identificador do hospital, sexo, data de nascimento" (Chu *et al.*, 1995), o processo de pesquisa dos mecanismos atualmente disponíveis não possuem bom nível de intuitividade e navegação de modo a auxiliar a exploração por itens

relacionados semanticamente, ou similares.

Figura 12 - Fontes de Conhecimento Físicas do CIT

Por se tratar de um contexto onde o atendimento emergencial é efetuado predominantemente via ligações telefônicas, a manipulação das pesquisas nos mecanismos computacionais torna-se dificultosa, pois em alguns casos, o sistema tem que ser operado com apenas uma mão, no caso de centros que não possuem aparelhos telefônicos com *headset*.

Na seção posterior será aprofundado o tema do Fluxo e Modo de Atendimento de urgência em Centros de Informações Toxicológicas – CIT, tomando como estudo de caso o CIT de Santa Catarina - CIT/SC.

3.3. O FLUXO DE ATENDIMENTO NO CENTRO DE INFORMAÇÕES TOXICOLÓGICAS

3.3.1. Como funciona o CIT

O Centro de Toxicologia de Santa Catarina (CIT/SC), formado por uma equipe de professores, médicos e farmacêuticos, além de acadêmicos de farmácia, medicina, biologia e ciências da informação, desenvolve seu trabalho integrando as três grandes áreas de extensão, ensino e pesquisa desde 1983. O CIT/SC atua na pesquisa epidemiológica e clínica, principalmente com as classes de animais peçonhentos, agrotóxicos e medicamentos (Cit, 2008). Presta também assistência direta ao paciente, ou auxilia o médico através do telefone nas emergências em todo o estado, além de desenvolver projetos de capacitação de recursos

humanos em toxicologia clínica e toxicovigilância para profissionais do SUS do estado de SC.

O atendimento no CIT pode ser efetuado de 2 formas, a saber: atendimento à distância (via telefone) e atendimento in loco. O atendimento in loco pode ser prestado a pacientes que se dirigem diretamente até o CIT, bem como pode ser oferecido a pacientes que tiveram encaminhamento através das emergências pediátrico ou adulto.

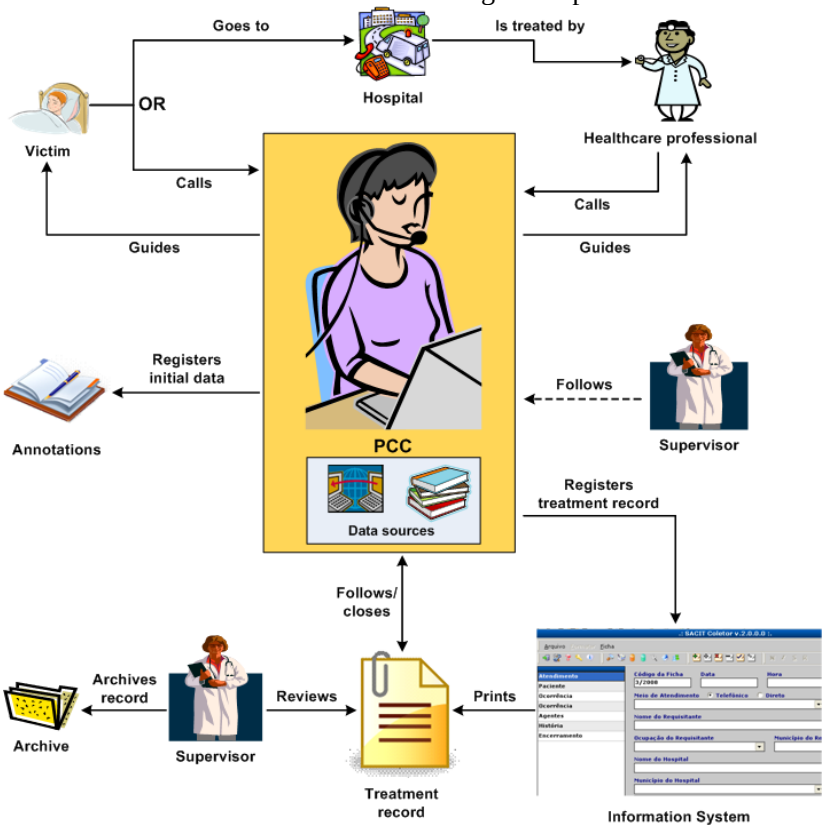


Figura 13 - Fluxo de trabalho em um atendimento efetuado no CIT de Santa Catarina
Fonte: (Ribeiro *et al.*, 2009)

O referido atendimento é realizado então por um plantonista – que pode ser um estagiário de medicina, farmácia ou áreas afins. No caso de atendimentos à distância, o plantonista se identifica e identifica o serviço (plantão CIT) ao telefone, aguardando a identificação do

paciente, anotando assim os principais dados cadastrais referentes ao mesmo.

Tendo isso sido feito, o plantonista começa então a inquirir o paciente com relação à intoxicação reclamada e toma nota dos detalhes que têm referencia com o ocorrido, efetuando uma espécie de anamnese.

Com a coleta básica de informações cadastrais em mãos, o plantonista começa a fazer a busca pelos dados nas bases de toxicologia para prestar o atendimento correto ao paciente. Para isso, ele tem à sua disposição um acervo de monografias com informações toxicológicas; um software desenvolvido por funcionários do CIT para auxílio ao atendimento, com informações de
diversos tipos de intoxicação, além do acesso ao site do Micromedex (site pago que possui uma grande quantidade de informações médico-farmacêutica e industrial).

Sempre com acompanhamento de um supervisor, o plantonista sugere tratamento para a intoxicação baseado nas informações pesquisadas por ambos. Assim, dando continuidade, é realizado o acompanhamento do tratamento do paciente através de ligações telefônicas ao próprio paciente ou médico que encaminhou o caso.

A ilustração do workflow da atividade de prestação de atendimento no CIT de Santa Catarina pode ser visualizada na Figura 13.

3.3.2. As fontes de dados/informação relacionadas ao contexto do CIT

Um dos objetivos deste trabalho é criar uma base de conhecimento no contexto de Toxicologia Clínica no sentido de homogeneizar e integrar várias fontes de conhecimento sobre este respeito. Desta forma, efetuou-se uma pesquisa sobre fontes potencialmente relevantes a este contexto, tendo sido examinadas as que serão mencionadas na sequência:

- BASE LOCAL (CIT/SC)

Os Centros de Informações Toxicológicas de Santa Catarina e do Rio Grande do Sul possuem suas próprias bases de dados sobre agentes toxicológicos, construídas de forma empírica, aplicando diferentes processos para consolidação de informações.

- LOINC

O Banco de dados LOINC - *Logical Observation Identifiers Names and Codes* tem por propósito facilitar o intercâmbio de resultados clínicos, fornecendo um conjunto de códigos universais e de laboratório para

identificar nomes e outras observações clínicas. Cada registro de LOINC corresponde a um único resultado de teste e inclui campos:

- componente (analyte) - por exemplo, potássio, hemoglobina, antígeno da hepatite C;
- propriedade medida - por exemplo, uma concentração maciça, atividade de enzima (taxa catalítica);
- a medida é uma observação em um momento do tempo, ou uma observação integrou sobre uma duração prolongada do tempo;
- o tipo de amostra - por exemplo, urina; sangue;
- o tipo de escala;
- o método que se usou para produzir o resultado ou a outra observação.

Observou-se que apenas uma fração do vocabulário LOINC está relacionada à toxicologia cujo objetivo é estabelecer uma terminologia comum para o intercambio de resultados clínicos, e portanto, não sendo útil para a construção da ontologia sobre os agentes toxicológicos que fará parte da base de conhecimento, pois não contempla a definição conceitual e as eventuais relações conceituais (de domínio ou de linguagem) necessárias para a composição da ontologia.

- RxNORM

RxNorm é tanto uma fonte como um subconjunto do Metathesaurus. O Escopo do RxNorm e determinado pela combinação do escopo de seus vocabulários fonte. Muitos relacionamentos (principalmente sinônimos), atributos de conceitos, e alguns nomes de conceitos são adicionados pela *National Library of Medicine* - NLM durante a criação das formas do RxNorm, mas essencialmente todos os conceitos vem de um ou mais vocabulários fonte.
No site do UMLS Knowledge Source Server está disponível para uso ou download o banco de dados do RxNorm em variadas formas. Para download, estão disponíveis versões dos dados para carga em um banco de dados MySQL ou Oracle.

- UMLS

A UMLS (Bodenreider e Burgun, 2004) é mais um projeto iniciado pela NLM em 1986, e empreende esforços para superar dois significantes

obstáculos para a efetiva recuperação de informação legível por computador:

- A grande variedade de meios e formas pelos quais os mesmos conceitos estão disponíveis e por diferentes pessoas;
- A distribuição de informação dispersa em várias bases de dados e sistemas.

Neste contexto, a UMLS objetiva se constituir um "*middleware* conceitual", um conjunto de ferramentas para desenvolvedores de sistemas destinados a manipular e utilizar conhecimento sobre a área médica. Por meio do tratamento e organização de conceitos de várias fontes (ex: MeSH (Nlm, 2006), RxNORM, LOINC e muitas outras), é possível acessar estes conceitos de uma forma padronizada e unificada. Alguns usos possíveis são a recuperação de informação na área médica, a construção de tesauros especializados (ex: doenças ou elementos sobre toxicologia), processamento de linguagem natural (PLN), indexação automática, mineração de dados e descoberta de conhecimento, acesso ao *Eletronic health records* (EHR), desenvolvimento de ontologias entre outros.

A UMLS é composta por três fontes de conhecimento: Metathesaurus, composto por mais de 1.000.000 de conceitos em 17 idiomas diferentes, mapeados na forma de um tesauro; Semantic Network, composta por uma estrutura similar a uma ontologia, agrupando 135 grandes categorias e 54 relacionamentos entre as categorias; SPECIALIST Lexicon & Tools, contendo informação e recursos para uso em PLN.

Os recursos da UMLS estão disponíveis no UMLS *Knowledge Source Server* (UMLS Server), e requer que o usuário solicite previamente uma licença de uso dos seus recursos, o que pode ser feito mediante cadastro. Neste local estão disponíveis recursos para navegação, download de dados e programas e documentação. O Metathesaurus é um banco de dados que contém um vocabulário multilíngüe, para múltipos fins e em larga escala que contém informações sobre conceitos relacionados às áreas da saúde e biomédicas, suas denominações e os relacionamentos entre eles, sendo que, sua constituição é elaborada a partir das versões eletrônicas de diferentes tesauros, classificações, conjuntos de códigos, e listas de termos controlados utilizados na assistência ao paciente, faturamento de serviços de saúde, estatísticas sobre saúde pública, indexação e catalogação da literatura biomédica, e/ou pesquisa básica,

clínica e em serviços de saúde. É organizado por conceito ou significado, associando e agrupando nomes alternativos e visões do mesmo conceito em conjunto para identificar as relações úteis entre diferentes conceitos.

- ANVISA

A base de medicamentos da ANVISA (Anvisa, 2009) contém uma relação com cerca de 20.000 medicamentos, contendo informações como a denominação, apresentação, princípios ativos (substâncias), dados do fabricante, entre outras. Sua obtenção para análise foi feita através do CIT/SC.

- Drugs FDA

O *U.S. Food and Drug Administration* – trata-se de um organismo do governo norte-americano que possui, entre outras atribuições, o controle do licenciamento de medicamentos comercializados naquele país. No site do FDA está disponível uma base de dados na forma de arquivos texto, prontos para importação para um banco de dados, com a relação de Drogas e Princípios ativos homologados pelo FDA. Através da análise feita sobre a relevância dos dados desta base, chegou-se a conclusão de que esta fonte não seria utilizada como insumo para a produção da ontologia em função de não conter medicamentos comercializados no Brasil, limitação resolvida a partir da obtenção da base de medicamentos da ANVISA.

- INTOX

International Programme on Chemical Safety: Banco de dados para consultas sobre informações em Toxicologia utilizadas nos CIT's. O download do software e do banco de dados é pago, mediante uma taxa anual, fixada na época deste trabalho em U$250 por ano.

- NTOX

O banco de dados INTOX é uma coleção de documentos úteis para aqueles que atuam em CIT's ou que está envolvido em gerenciamento e diagnóstico sobre intoxicação. O conteúdo inclui documentos revisados por pares - em nível internacional - sobre produtos químicos, farmacêuticos, agropecuários e toxinas de plantas, fungos e animais, e sobre o tratamento de intoxicações, toxicologia analítica e operações em CIT's. Muitos destes documentos, como monografias sobre intoxicações e guias de tratamento que são escritas por participantes do IPCS INTOX Programme.

- INTOX DATA MANAGEMENT SYSTEM

O Sistema de Gerenciamento de Dados INTOX disponibiliza aos CIT's acesso a informações em bancos de dados para pesquisa, substâncias e produtos. Cada um destes bancos de dados tem as informações organizadas de forma consolidada, permitindo rápida recuperação por parte dos CIT's. O sistema é disponível em Inglês, Francês, Português e Espanhol, usando vocabulários controlados para o tratamento documentário.

- HSDB

HSDB (Hsdb, 1991) é um banco de dados sobre toxicologia da NLM, *Toxicology Data Network* (TOXNET) e contém informações sobre a toxicologia dos produtos químicos potencialmente perigosos. Contém informações organizadas nos seguintes grupos:

- efeitos sobre saúde humana;
- tratamento médico de emergência;
- estudos sobre toxicidade em animais;
- farmacocinética;
- farmacologia;
- risco de exposição ambiental e outras informações correlatas.

Todos os dados são referenciados, e obtidos a partir de livros, documentos governamentais, relatórios técnicos e periódicos primários previamente selecionados. Os registros do HSDB são revisados pelo Painel Científico Review (SRP), uma comissão de especialistas nas principais áreas sujeitas a banco de dados dentro desta área.

- ANVISA – Lista DCB

O Brasil dispõe de lista de Denominações Comuns Brasileiras (DCB), periodicamente atualizadas, apresentando cerca de 9.370 denominações genéricas, de propriedade pública e oficial, utilizadas em dossiês de registros de medicamentos, licitações, manipulação de medicamentos, rastreamento de insumos, prescrição médica, legislação e qualquer forma de trabalho ou pesquisa científica.
Esta lista referencia a nomenclatura de substâncias utilizadas nos medicamentos no Brasil, composta pelo nome da substância (em português) e o respectivo registro CAS.

- DeCS

O vocabulário DeCS (Descritores em Ciências da Saúde) foi criado pela BIREME (Bireme, 2010) em 1982 para uso na indexação de assuntos e na recuperação em bases de dados da área da saúde. É uma extensão no MeSH - *Medical Subject Headings* da U.S. *National Library of Medicine* (Nlm, 2006) oferecendo uma terminologia em português, inglês e espanhol, para a recuperação da informação independentemente do idioma. Possui 30369 descritores na versão 2010, agrupados em 20 categorias, entre as quais, 04 são desenvolvidas no Brasil nas áreas de Saúde Pública, Homeopatia, Ciência e Saúde e Vigilância Sanitária.

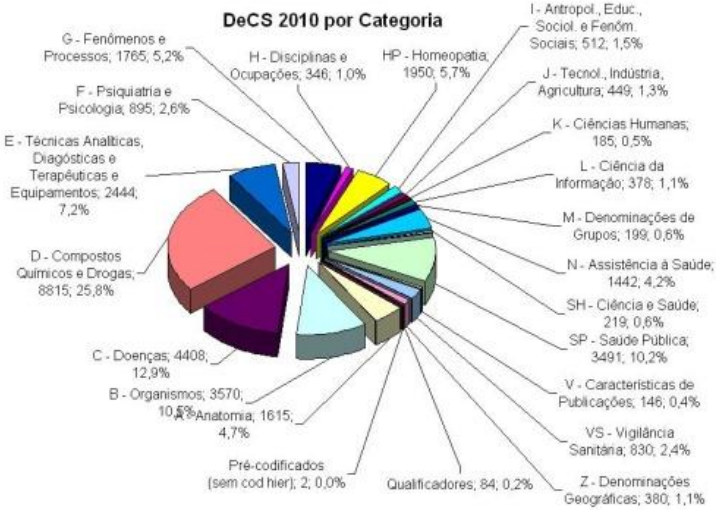


Figura 14 - Distribuição dos descritores nas Categorias do DeCS
Fonte: (Bireme, 2010)

- AGROFIT

A base de dados AGROFIT (Agrofit, 2010) contém o cadastro de produtos agrotóxicos e afins registrados no Ministério da Agricultura, Pecuária e Abastecimento, disponível para consulta ao público em geral, com informações detalhadas sobre estes produtos, inclusive com informações sobre sua toxicidade.

3.4. CENÁRIO COMPARATIVO: A ATIVIDADE DE PESQUISA DE AGENTES – ANTES E DEPOIS

Como já visto anteriormente, a presente proposta visa promover maior

conforto ao usuário e melhor performance no momento das buscas por agentes.

O atual processo de pesquisa de agentes é ilustrada na Figura 15.

Através da ilustração é possível observar o *workflow* que o atendente precisa seguir para realizar uma pesquisa. Considerando a pior das hipóteses, pode ser necessário pesquisar nas diversas fontes de conhecimento disponíveis para esta pesquisa, havendo a possibilidade de encontrar o resultado esperado somente na busca efetuada na última tentativa/fonte, percurso este que é demarcado pela linha pontilhada na Figura 15.

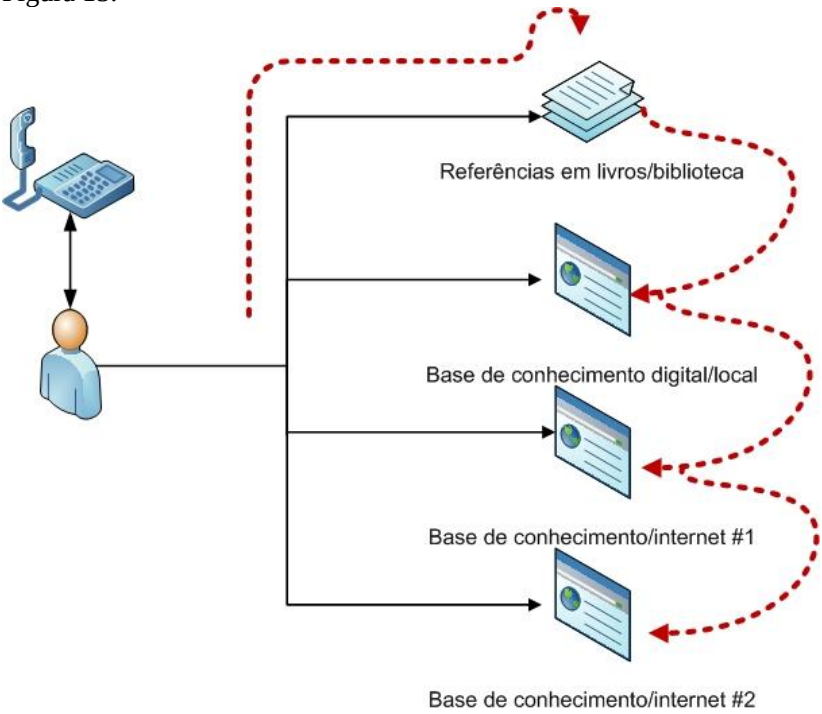


Figura 15 - Cenário da atividade de pesquisa por agentes tóxicos sem a utilização de um motor de busca integrado

A motivação de se implementar uma solução integrada de busca tem a visão de economizar tempo e esforço do atendente, possibilitando que esta pesquisa seja efetuada em apenas uma interface, onde é efetuado o retorno desta pesquisa. A redução do percurso do workflow pode ser observada na linha pontilhada, da Figura 16.

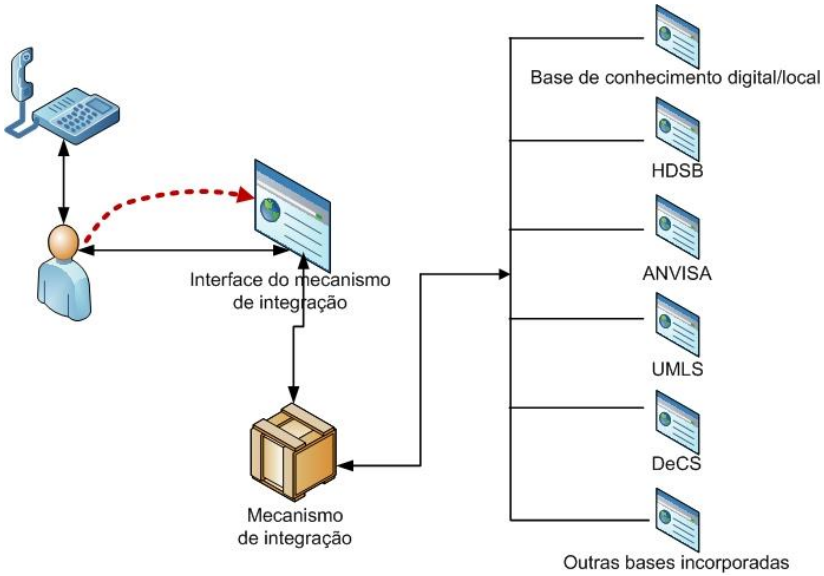


Figura 16 - Cenário da atividade de pesquisa por agentes tóxicos com a utilização de um motor de busca integrado

3.5. ENGENHARIA DO CONHECIMENTO

Nesta seção, será feita uma breve abordagem sobre a gênese da Engenharia do Conhecimento, trazendo conceitos da Inteligência Artificial como ferramentas de produtividade e aplicação em Sistemas de Conhecimento.

Por meio de uma breve revisão de materiais e métodos utilizados na IA, será possível efetuar a correlação com a atual abordagem de Sistemas de Conhecimento que é proposto neste trabalho.

3.5.1. Inteligência Artificial

Conforme argumentam os autores (Wielinga *et al.*, 1997; Abel, 2002), a aquisição e processamento do conhecimento têm sido, tipicamente, atividades pesquisadas na área de Inteligência Artificial. Desta forma, torna-se importante uma revisão de alguns conceitos básicos da IA que foram recebidos como herança de alguma maneira pelos Sistemas de Conhecimento.

Abel explica ainda que as atividades de manipulação do

conhecimento podem ser visualizadas principalmente no âmbito do estudo do comportamento humano e a reprodução por similaridade deste comportamento pelo computador, através de um conjunto de programas e sua arquitetura em IA. Este estudo divide a Inteligência Artificial em três grandes áreas:

- Processamento de linguagem natural – PLN - trata-se da área que tem como objetivo facilitar o uso dos computadores permitindo-os a comunicação com os usuários através de linguagem natural (Paris *et al.*, 1991);
- Robótica, que como o próprio nome já diz, trata de funções de movimento, percepção e controle através de hardware e software com algoritmos específicos, como pode ser visto em (Russell e Norvig, 2009);
- Processamento de Conhecimento, que reflete as características de armazenamento, manipulação e reuso do conhecimento, tendo em vista como uma das utilizações possíveis o auxílio à resolução de problemas, como pode ser visto detalhadamente em (Clancey, 1985; Hansen *et al.*, 2005).

Uma ilustração do paralelo entre funções e as áreas da Inteligência Artificial, conforme pode ser visto na Figura 17.

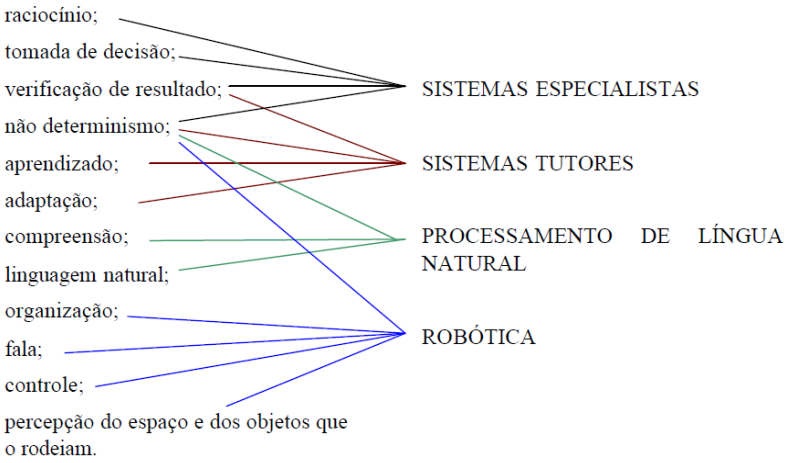


Figura 17 - Representação do relacionamento entre funções e as áreas da IA
Fonte: (Abel, 2002)

3.5.2. Engenharia do Conhecimento: o surgimento da disciplina

O termo "Engenharia do Conhecimento" foi cunhado na década de 80 em referência aos processos de elicitação do conhecimento retido por especialistas em determinadas áreas, no intuito de, a partir desta elicitação, construir sistemas baseados em conhecimento (Shaw e Gaines, 1992), os quais serão vistos mais adiante.

Historicamente, o surgimento da disciplina de Engenharia do Conhecimento tem seu nascimento através dos primeiros experimentos com desenvolvimento de mecanismos e ferramentas que operacionalizaram os primeiros Sistemas Baseados em Conhecimento (SBCs) que foram realizados na disciplina de Inteligência Artificial, a fim de estudar a viabilidade de seu uso. Tipicamente, estes experimentos eram restringidos a pequenas aplicações com o intuito unicamente de avaliar tal viabilidade (Studer *et al.*, 1998).

Segundo Studer, no momento da transferência desta tecnologia para ambiente comercial com o intuito de desenvolver grandes aplicações baseadas em conhecimento, ela (tecnologia) se mostrou falha em muitos casos, o que gerou uma situação em que compara com a "crise do software", ocorrida nos anos 60 devido aos *softwares* produzidos na academia não terem bom desempenho quando adaptados a contextos âmbitos comerciais.

Em ambos os casos, havia a necessidade de disciplinas que tornassem o processo de desenvolvimento uma arte fundamentada sobre uma disciplina de engenharia, demandando análise e processos de manutenção através de métodos, linguagens e ferramentas apropriados para tal e, no caso da manipulação do conhecimento e através desta demanda é que surge a disciplina de estudo em questão.

3.5.3. Conhecimento como fator de valor agregado (Capital Intelectual)

"*Conhecimento é poder"* (Francis Bacon *apud* Liao, 2003). Esta é uma afirmação que por si só caracteriza a importância do artefato tratado neste trabalho, que tem por objetivo "*preservar valores, aprender novas coisas, resolver problemas, criar competências e iniciar novas situações tanto individual quanto organizacionais, agora e no futuro*" (Liao, 2003).

Tom Stewart (1997) comenta que após a "*Era Industrial*" o Conhecimento passou a ser um importante fator de produção, mesmo

que algumas vezes não seja uma característica aparente no que diz respeito ao lucro de uma empresa. A dificuldade desta caracterização diz respeito a sua forma ou tangibilidade (Bogdanowicz e Bailey, 2003). A importância de poder manipular o conhecimento nos dias atuais têm sido alvo dos mais diversos estudos, com o intuito de torná-lo um elemento de lucro nas empresas. Stewart exemplifica também mais uma aplicação do conhecimento com valor agregado:

> "A matéria prima essencial da Revolução Industrial era óleo e aço. Bem, mais de 50% do custo de extrair petróleo da terra atualmente é obter e processar informação. Assim como para o aço, grandes produtores precisavam de quatro horas-homem de trabalho para fazer uma tonelada de aço. Agora, com o uso de sofisticados computadores, precisam apenas de 45 minutos de trabalho por tonelada. O componente intelectual cresceu e o físico encolheu.
>
> Se o aço foi o produto essencial da industrialização, o produto da Idade da Informação é o circuito integrado (CI). O valor de todos os CI's produzidos excede o valor do aço produzido. O que faz com que tenham esse valor? Certamente não são seus componentes físicos. Um CI é feito basicamente de silício, ou seja, de areia e em pouca quantidade. O valor está principalmente no projeto do circuito, e no projeto das complexas máquinas que o fazem. Esse ingrediente principal é conhecimento.
>
> A soma disto tudo nos leva a uma conclusão: mais e mais do que nós compramos e vendemos é conhecimento. Conhecimento é a principal matéria prima" (Stewart, 1997).

Surgem a partir daí diversas abordagens sobre "trabalhadores e trabalho do Conhecimento" (Boff, 2000) e o tratamento do Conhecimento como um *commodity* (Van Der Spek e Spijkervet, 1997), enfoques que auxiliam na caracterização do Conhecimento como **Capital Intelectual**.

3.5.4. Evolução da Engenharia do Conhecimento: mudança de paradigma

Uma das características da disciplina de Engenharia do Conhecimento é a diferenciação no que diz respeito na sua capacidade evolutiva, que pode ser vista em uma análise do acompanhamento das tendências de

sua aplicação. Esta caracterização é remetida historicamente no evento da troca de paradigma que será abordada na sequência.
Segundo (Studer *et al.*, 1998), a Engenharia do Conhecimento é dividida em dois períodos, que foram chamados de "*abordagem de transferência*" e "*abordagem de modelagem*", que para alguns pesquisadores foi considerada a transição entre a primeira e segunda gerações dos sistemas especialistas (David *et al.*, 1993).

- **Abordagem de transferência**

Musen (1993) relata que o desenvolvimento na década de 80 era visto como um *processo de transferência* do conhecimento humano implementado em uma base de conhecimento, e baseava-se na afirmação de que o conhecimento requerido para os Sistemas Baseados em Conhecimento (SBCs) já era existente, e desta forma, bastava ser coletado e implementado. A partir daí, o processo de desenvolvimento teve mais ênfase na para as técnicas de extração do conhecimento de especialistas e a codificação e formalização deste conhecimento (Abel, 2002). Dentre os sistemas que adotam esta arquitetura, podemos citar o MYCYN (Davis *et al.*, 1977) e PROSPECTOR (Hart *et al.*, 1978).

Todavia, após fazer um estudo cuidadoso sobre a ferramenta MYCYN, Clancey (1983) exemplifica através dela a problemática envolvida em uma grande parte dos sistemas baseados em conhecimento desenvolvidos na época: havia uma grande quantidade de regras contidas em diferentes tipos de domínios e, a mistura destes tipos de conhecimento aliada à falta de justificativas adequadas para estas regras se tornava, segundo Clancey, fator agravante para a dificuldade na manutenção destas bases de conhecimento.

O reconhecimento das falhas que pelas quais este paradigma foi acometido foi ponto de partida para a mudança de visão na engenharia de sistemas de conhecimento, aliada a consideração de que nem todo o conhecimento necessário para a resolução de problemas especializados já é existente, ressaltando a importância do conhecimento tácito adquirido pelos especialistas em resolução dos problemas referidos (Wielinga *et al.*, 1997; Studer *et al.*, 1998; Schreiber, 2000).

- **Abordagem de modelagem**

Na atualidade, existe um consenso de que o processo de construção de SBC's pode ser considerado uma atividade de modelagem (Studer *et al.*, 1998). O foco considerado está na construção de sistemas que possam operar na resolução de problemas com capacidade comparável à de um especialista de domínio, sem a intenção de produzir sistemas com

aspecto cognitivo, mas produzir resultados semelhantes a esta operação se fosse realizada pelo referido especialista.

Fase	Nome	Característica	Exemplos
1ª.	Sistemas Simbólicos	Conhecimento extraído de um ou mais indivíduos de modo empírico. Sistemas de processamento simbólico. Regras embutidas no código do programa.	Sistema Macsima Sistema Dendral Sistema Puff
2ª.	Sistemas Especialistas	Conhecimento extraído de um indivíduo com técnicas de aquisição de conhecimento. Base de conhecimento separada do mecanismo de inferência.	Sistema Mycin Sistema Propector
3ª.	Engenharia de Conhecimento	Conhecimento organizacional, racionalizado a partir da experiência de muitos e padronizado de acordo com os objetivos da organização. Método de inferência também modelado na base de conhecimento. Base de conhecimento pode estar distribuída e o sistema é integrado aos demais da organização.	Metodologia CommonKADS Metodologia Protege Metodologia Vital

Figura 18 - Evolução da Engenharia do Conhecimento
Fonte: (Abel, 2002)

O que difere principalmente da abordagem de transferência vista anteriormente é a consideração das habilidades tácitas, que devem ser adquiridas em um processo de aquisição de conhecimento, e que necessariamente é considerada uma atividade de modelagem (Clancey, 1989; Morik, 1991; Studer *et al.*, 1998), e caracteriza uma visão de modelagem de construção com as seguintes características e conseqüências:

- Como todo modelo, o modelo é apenas uma aproximação da realidade;
- A modelagem é um processo cíclico, ou seja, novas observações são feitas a cada refinamento, cada ciclo;

- O processo de modelagem está sujeito às interpretações do engenheiro do conhecimento.

Remete-se através desta evolução de visão ocorrida na Engenharia do Conhecimento o estudo de novas metodologias que pudessem atender aos requisitos necessários para modelar "*processos de solução de problemas que fossem racionalizados e padronizados por uma organização, e não apenas a reprodução do conhecimento de um especialista*" (Abel, 2002).

A Figura 18, de autoria da professora Mara Abel, ilustra pragmaticamente um histórico da evolução da Engenharia do Conhecimento segundo o aspecto de desenvolvimento, elucidando os acontecimentos, enfoques e artefatos produzidos do período na história contextualizada, as características do tipo de enfoque utilizado na época de aplicação e os exemplos mais amplamente divulgados de utilização de cada abordagem referenciada durante esta evolução.

3.5.5. Sistemas de Conhecimento

Schreiber postula que os Sistemas de Conhecimento fossem vistos como ferramentas para Engenharia do Conhecimento, devido ao oferecimento de soluções em potencial para problemas de conhecimento dectectados, analisados e priorizados pela Engenharia do Conhecimento (Schreiber, 2000).

A comunidade da Engenharia do Conhecimento tem, ao longo dos anos, desenvolvido uma gama muito grande de métodos e técnicas para aquisição, modelagem, representação e reuso do conhecimento (David *et al.*, 1993; Schreiber *et al.*, 1994; Wielinga *et al.*, 1997), com o propósito de construir sistemas baseados em conhecimento que executassem tarefas intensivas em conhecimento.

Através da introdução do novo paradigma da Engenharia do Conhecimento surgiram também metodologias para auxiliar no desenvolvimento de sistemas de conhecimento. Dentre estas metodologias podemos citar:

- MIKE (Angele *et al.*, 1998);
- VITAL (Nigel, 1993);
- COMMET (Steels, 1993);
- EXPECT (Swartout e Gil, 1995);

Para ilustrar a estrutura de uma metodologia, na seção **Métodos e Técnicas em Engenharia do Conhecimento** será descrita como exemplo a metodologia **CommonKADS** conforme (Wielinga *et al.*, 1997), utilizada conceitualmente para o desenvolvimento deste trabalho.

3.5.6. Ontologias

Com o aumento exponencial da produção de conhecimento em virtude do incentivo à pesquisa em diversas áreas da ciência, descobriu-se uma demanda de organização deste conhecimento, a qual se atribui uma importância significativa.

A partir desta demanda, para obter organização da informação foram criadas técnicas que fazem parte de um corpo de disciplinas que visam obter um melhoramento no tratamento de dados, atuando sobre sua classificação, processamento, recuperação e disseminação (Guimarães, 2002; Almeida e Bax, 2003).

Embora atualmente o termo Ontologia seja amplamente utilizado no contexto tecnológico, este não tem origem nesta área, como pode ser visto a seguir:

- **Ontologia na Filosofia**

Na filosofia, a ontologia trata do "*estudo das coisas que existem*" (Chandrasekaran *et al.*, 1999).

No contexto histórico, o termo ontologia foi cunhado originalmente no grego, pela concatenação dos vocábulos "onto" e "logos", traduzidos para "ser" e "palavra", respectivamente.

Origina-se da palavra Aristotélica "categoria", utilizada para classificar elementos. Aristóteles utilizava as categorias como base para classificar quaisquer entidades e, desta forma, possibilitando diferenciar espécies de um mesmo gênero.

Ao tratar do estudo filosófico de ontologia, é aplicado propriamente como sendo a distinção do "Estudo do Ser" (Almeida e Bax, 2003).

Em sistemas tecnológicos, ontologia tem uma conotação pouco semelhante à utilizada na filosofia, como abordado no item seguinte.

- **Ontologias em Sistemas de Conhecimento**

Entre as primeiras definições para ontologias utilizadas no contexto tecnológico está a de (Gruber, T., 1995), onde declara que "uma ontologia é uma especificação explícita de uma conceitualização". Willem Borst determinando aquela que atualmente é uma das mais utilizadas definições para ontologia, onde descreve ontologia como

sendo “uma especificação explícita e formal de uma conceitualização compartilhada” (Borst, 1997).

Em Ciências da Computação, este conceito é apropriado para descrever um artefato resultante de processos de Engenharia do Conhecimento e é definido por Sowa (1999) como “*um catálogo de qualquer coisa que constitui um mundo, como são colocadas e como funcionam*”.

Em síntese, o termo foi adaptado para a comunidade de inteligência artificial como referência a um conjunto de conceitos usados para descrever um domínio (Rios, 2005).

Guarino (1998) afirma que a palavra “ontologia” é o nome que é utilizado para definir as atividades de analise conceitual e modelagem de domínio, suportado por metodologias advindas de outros campos das Ciências da Computação. Studer et. al. (1998) faz uma observação sobre uma confusão rotineira relacionada ao entendimento sobre o que vem a ser ontologia, afirmando que este conceito muitas vezes é confundido com outros tipos de representações, como taxonomias, por exemplo.

A partir desta observação realizada no estudo de publicações relacionadas a ontologias, Guarino argumenta que a construção de ontologias deve ser um processo interdisciplinar, em que a filosofia e a lingüística possam fornecer as regras fundamentais para a análise da estrutura do domínio com alto nível de clareza da linguagem e expressividade.

A partir destas definições, é possível compreender a relação do conceito de ontologia e seus componentes, que serão descritos na sequência.

- **Características de uma Ontologia (Componentes Básicos)**

Segundo (Novello, 2002):

> “- Conceitos representam qualquer coisa do domínio sobre a qual alguma coisa é dita; incluem os objetos do domínio, a descrição de uma tarefa, de uma função, ação, estratégia etc;
> - Relações representam os tipos de interações entre os conceitos do domínio. São definidas formalmente como qualquer subconjunto de um produto de n conjuntos, ou seja: R:C1 x C2 x ,..., x Cn;
> - Funções são relações especiais onde o n-ésimo elemento da relação é único para os n-1 elementos precedentes; formalmente, funções são definidas como F: C1 x C2 x ,..., x Cn-1 → Cn; são

> exemplos de funções antecedente-de e causa, indicando que o valor do segundo componente da relação depende do primeiro;
> - Axiomas modelam sentenças que são sempre verdadeiras são classificados em estruturais e não estruturais;
> - Instâncias representam elementos da ontologia" (Novello, 2002).

- **Para que serve uma Ontologia?**

Chandrasekaran (et al. 1999) afirma que as ontologias contribuem para elucidação da estrutura do conhecimento. Ontologias podem ser tomadas como modelo para projetar domínios de gestão do conhecimento, comércio eletrônico, processamento de linguagens naturais, recuperação de informação na web, informações educacionais, etc. (Almeida e Bax, 2003).

Ainda segundo Chandrasekaran, dado um domínio, é a ontologia que se torna o coração de qualquer sistema de representação do conhecimento deste domínio, e sem elas, não é possível que um vocabulário represente este conhecimento. Outro ponto de extrema importância relacionado a ontologias é a capacidade de compartilhar conhecimento, além de viabilizar a padronização, interoperabilidade, recuperação e reuso da informação e resolver problemas de confusões terminológicas e troca de informações entre softwares, entre outros (Manica *et al.*, 2009).

- **Metodologias para construção de Ontologias**

A atividade de desenvolvimento de uma ontologia demanda competências e recursos para efetivar, de forma satisfatória, o processo de representação do conhecimento do domínio. Desta forma, faz-se necessária uma abordagem sistemática do domínio para compreendê-lo e torná-lo explícito, para que no escopo de aplicação da ontologia, sejam acolhidas as demandas de conhecimento para as quais foi projetada.

De acordo com Gruber (1995), Ontologias formais são projetadas. Quando se escolhe a maneira de representar algo em uma ontologia, se está definindo algo que irá interferir na sua qualidade. Para esta definição, é necessário basear-se em critérios devem ser orientados por algumas diretrizes, a saber:

a) Clareza: a ontologia deve, com efetividade e objetividade, os significados nela definidos;

b) Coerência: os conceitos declarados devem ser consistentes, bem como a estrutura lógica dos seus axiomas;

c) Extensibilidade: o projeto de uma ontologia deve antecipar a possibilidade de expansão e reuso por meio do seu compartilhamento;

d) Mínimo viés de codificação: a conceitualização deve ser especificada no nível do conhecimento, sem depender de um padrão particular de codificação;

e) Mínimo compromisso ontológico: a ontologia deve possuir um compromisso ontológico com o objeto representado o suficiente para que ela seja compartilhada e reusada.

Para auxiliar no atendimento dos requisitos previamente descritos são definidas metodologias para o desenvolvimento de ontologias. Algumas delas são brevemente descritas abaixo:

- **Cyc** - desenvolvida em meados de 1980, pela Microeletronics and Computer Technology Corporation (MCC). Nesta metodologia, todo o conhecimento é representado declarativamente sob a forma de afirmações em uma variante da lógica de primeira ordem chamado CICL. A base de conhecimento CYC por si mesma contém afirmações simples, regras de inferência e regras de controle para inferências; um mecanismo de inferência pode ser usado para derivar novos argumentos usando esta base de conhecimento (Uschold e Gruninger, 1996).

- **Uschold and King's** – em 1995, primeiro método proposto para construção de ontologias propriamente dito, onde os autores definem uma lista de etapas a serem seguidas para a construção de ontologias, os quais, segundo os autores, poderiam ser úteis para quaisquer metodologias que fossem propostas no futuro (Uschold e King, 1995);

- **Grüninger and Fox's** – metodologia baseada na experiência dos autores na construção de ontologias (Gruninger e Fox, 1995);

- **KACTUS** – proposta em 1996 por Amaya Bernaras e seus colegas, dentro do projeto Esprit KACTUS (Schreiber *et al.*, 1995);

- **METHONTOLOGY** – desenvolvida no grupo de ontologia da Universidad Politécnica de Madrid, permite a construção de ontologias no nível de conhecimento (Fernandez *et al.*, 1997).

3.5.7. Métodos e Técnicas em Engenharia do Conhecimento

Nesta subseção serão discutidos métodos e técnicas úteis para a construção de Sistemas de Conhecimento. Como visto anteriormente, metodologias são itens fundamentais para a construção de um sistema de Conhecimento consistente. No inicio desta seção, serão demonstrados brevemente conceitos relacionados à metodologia CommonKADS, seguindo pela ilustração de técnicas de Recuperação da Informação, técnicas de Expansão de Consulta e, finalmente, técnicas de Anotações Semânticas.

- **CommonKADS**

Trata-se de uma metodologia que, segundo seu idealizador (Schreiber, 2000), originou-se da necessidade de se criar sistemas de conhecimento em larga escala de qualidade industrial, de maneira estruturada, controlável, e repetida.

O CommonKADS tem sua gênese na metodologia KADS, que objetiva principalmente o suporte ao desenvolvimento de sistemas baseados em conhecimento, dando ênfase não só à aquisição e representação do conhecimento especializado mas também o aspecto organizacional no âmbito de aplicação (Wielinga *et al.*, 1992). Em sua estrutura, há uma separação do conhecimento da aplicação e do conhecimento utilizado na solução de problemas (Abel, 2002). Embora sejam definidos nas mesmas primícias, um deles refere-se aos objetos do mundo real enquanto que o outro possui um objeto próprio de conhecimento da aplicação.

Um modelo de conhecimento CommonKADS possui uma série de tipos de conhecimento. Estes tipos de conhecimento têm diferentes papéis na resolução de problemas (Wielinga *et al.*, 1997).

No modelo de conhecimento referido se destacam cinco tipos de conhecimento (Schreiber, 2000):

  - Tarefa: especifica o objetivo em um modo funcional, indicando as entradas e saídas de uma

tarefa e as dependências lógicas entre estas entradas/saídas. Exemplo de uma tarefa: diagnóstico. A entrada pode ser uma reclamação sobre um sistema e a saída uma categorização da falha (Schreiber, 2000; Abel, 2002);

- Método: é a prescrição da solução de um problema ou subproblema. Como exemplo o autor cita geração e teste, proposta e revisão, etc. Existem dois tipos de métodos: genéricos, que são descrições independentes de uma tarefa em particular e, de tarefa, onde os métodos são aplicados para tarefas em particular. Para maiores informações sobre os tipos de métodos, é indicada a consulta em (Schreiber, 2000);
- Inferência: conhecimento de inferência descreve as inferências básicas que se deseja fazer sobre um domínio. Uma inferência opera em uma determinada entrada e tem a capacidade de produzir um novo artefato de informação como saída. Inferências podem operar sobre elementos de conhecimento, e estas elementos são descritos como papéis de conhecimento dinâmicos. Para efetuar estas operações, as inferências também são utilizadas em elementos de conhecimento que não são afetados por esta operação. Estes elementos são descritos como papéis de conhecimento estático (Schreiber, 2000; Abel, 2002);
- Esquema do conhecimento do domínio: são descritores esquemáticos da estrutura do domínio do conhecimento utilizado para a resolução de um problema. Um esquema de domínio do conhecimento tem uma função similar a do modelo de dados na engenharia de software tradicional, mas é mais complexa devido ao fato de conter *metaconhecimento* (conhecimento sobre o conhecimento). O termo "ontologia" é muito utilizado como sinônimo para o esquema de domínio de conhecimento;

- Conhecimento do domínio: trata-se de uma coleção de conceitos, relações e fatos que são utilizados para inferências (Schreiber, 2000; Abel, 2002).

Segundo Fensel e Harmelen (1994), para cada tipo de conhecimento são utilizadas linguagens e notações apropriadas para sua descrição. Dentro da metodologia CommonKADS foi desenvolvida uma linguagem para modelagem para permitir a descrição semi-formal do modelo de conhecimento (Schreiber *et al.*, 1994). Nesta linguagem existe um mecanismo de mapeamento para permitir a construção de ontologias multi-camadas em que a camada mais exterior representa tipos de conhecimento mais abstratos. Uma maior atenção foi dada às pesquisas para as técnicas de reuso devido à constatação de que o processo de Engenharia do Conhecimento é oneroso.

Para as tarefas de resolução de problemas, uma tipologia de tarefas foi desenvolvida baseada nas características do problema que as tarefas teriam finalidade de resolver. Tarefas de diagnóstico, *design*, configuração e agendamento são exemplos destas tarefas. os tipos de tarefas são ordenadas de maneira hierárquica, (como por exemplo, a tarefa de configuração é uma sub-tarefa de *design)* (Wielinga *et al.*, 1997; Schreiber, 2000).

Para alguns tipos particulares de tarefa, foram criadas pela comunidade de Engenharia do Conhecimento bibliotecas de métodos, como a *CommonKADS library* (Breuker e Van De Velde, 1994) com uma grande lista de implementações.

Abel (2002) explana que "*o projeto de um sistema de conhecimento é essencialmente um projeto de preservação da estrutura*". Assim sendo, a preservação da informação é um dos objetivos básicos na modelagem, desta forma a reusabilidade faz-se necessária, pois "*a relação entre os construtos de comunicação e de conhecimento coletados na fase de análise devem ser facilmente recuperados nos componentes da arquitetura*".

Wielinga *et al*. (1997) argumenta que esta relação é um pouco mais complicada, e explica:

> "It is an empirical fact that the structure of domain knowledge is partially determined by the way it is used in the problem-solving process ('the relative interaction principle'). This dependency hampers re-use, because it means that schema re-use depends on the task and/or method context for

> which it was first developed" (Wielinga *et al.*, 1997, p.74).

A solução apresentada neste contexto é a "possibilidade de definir esquemas de diferentes níveis de especificidades e conectá-los através de mapeamentos parciais. Este nivelamento do domínio de conhecimento pode ser utilizado para identificar diferentes tipos de domínios de conhecimento com 'status de reusabilidade' diferentes" (Wielinga *et al.*, 1997).

Os níveis de especificação considerados por estes autores são:

1) O esquema de aplicação é mais específico que o esquema conhecimento de domínio. Ele contém exatamente os tipos de domínio que são demandados a dada aplicação;

2) Um esquema específico para um método introduz a representação para um certo (ou conjunto de) método(s);

3) Um esquema para uma tarefa específica contém as conceitualizações inerentes à tarefa. Por exemplo, em uma tarefa de desenvolvimento as noções do comportamento e regras necessariamente precisam estar presentes;

4) Um esquema de domínio específico descreve todas as conceitualizações do domínio que são independentes da tarefa ou método empregado. Por exemplo, uma descrição estrutural de um dispositivo que é utilizado tanto nas tarefas de *design* quanto de diagnóstico;

5) O esquema de inter-domínio generaliza os domínios e provê descrições genérias que podem ser encontradas em classes de domínios. Por exemplo, em domínios técnicos de processos industriais, tipos de conhecimento de domínio reutilizáveis podem ser identificados;

6) Um esquema genérico descreve uma série de definições que devem ser mais ou menos universalmente verdadeiras. Estes esquemas são semelhantes às categorias Aristotélicas. A principal diferença é que na área de pesquisa da Engenharia do Conhecimento nenhum descrédito é dado no caso da incompletude de um esquema. Seu propósito é mais pragmático do que filosófico: o esquema deve poder ajudar a possibilitar o reuso das primeiras experiências na especificação do conhecimento.

São notáveis os trabalhos direcionados à criação de bibliotecas

e definições de esquemas, entre estes, pode-se citar os projetos KACTUS e Sisiphus (Schreiber *et al.*, 1995; Schreiber e Birmingham, 1996).

- **Recuperação de Informação**

Segundo Baeza-Yates e Ribeiro-Neto (1999), "*Recuperação da Informação (IR) é uma área da computação que trabalha com a representação, armazenamento, organização e acesso à itens de informação"*. Estes mesmos autores afirmam que a IR deve prover, através da representação da informação de forma mais organizada, uma forma de fácil acesso à informação desejada ao usuário.

As operações relativas aos processos de recuperação da informação dizem respeito à indexação, normalização e transformação dos termos em vetores. Anteriormente aos documentos serem retornados ao usuário, estes são classificados de acordo com um índice de relevância específico, definido por um algoritmo apropriado (Van Rijsbergen, 1986).

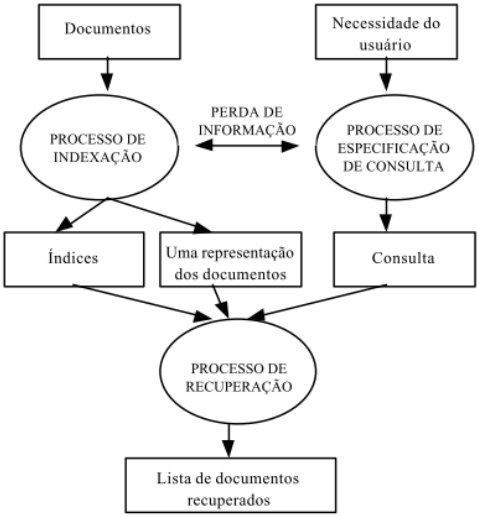


Figura 19 - Componentes de um sistema de recuperação de informação
Fonte: (Gey *apud* Cardoso, 2000)

Os primeiros sistemas de Recuperação de Informação utilizavam técnicas booleanas de recuperação. Com o início da sua utilização, foram demandadas combinações mais complexas de técnicas utilizando combinações de operadores booleanos, como por exemplo, os operadores *AND* (E), *OR* (OU) e *NOT* (NÃO), que acabavam por dificultar a formação de consultas que trouxessem bons resultados. Este

método é efetivo para recuperação de documentos baseados em palavras simples, mas é muito mais complexo para alguns tipos de consultas que poderiam levar muito tempo até executarem todas as iterações com o usuário (Manning *et al.*, 2008).

A fim de resolver estes problemas, novas técnicas foram desenvolvidas para aperfeiçoar os modelos tradicionais de Recuperação da Informação. Estas técnicas utilizam semântica para recuperar a informação, como proximidade de termos, uso de dicionários semânticos e léxicos, além de técnicas de PLN (Processamento de Linguagem Natural) para fazer a busca (Greengrass, 2001).

Na Figura 19 é possível identificar alguns dos componentes necessários a uma aplicação de Recuperação da Informação em uma estrutura padrão, conforme descrita por (Gey *apud* Cardoso, 2000).

- **Expansão de Consulta**

As técnicas de Expansão de Consulta foram propostas para aperfeiçoar o modelo de IR tradicional, no sentido de amenizar o problema de retorno de documentos não desejados a partir de uma consulta, concatenando termos adicionais e redefinindo pesos para estes termos, para melhorar seu desempenho (Cai *et al.*, 2001; Chang *et al.*, 2007).

Existem trabalhos sobre expansão de consulta que propõem a utilização de técnicas que pretendem avaliar o comportamento do usuário a fim de utilizá-lo como parâmetro na construção de novas consultas. As técnicas utilizadas para este fim são chamadas de *relevance feedback* (RF) e *pseudo-relevance feedback* (PRF), esta última, têm os termos gerados através da composição dos resultados mais bem classificados no retorno de busca (Yoo e Choi, 2010). Embora o primeiro método, que utiliza a intervenção do usuário para composição de modelos para novas consultas, o método automatizado – PRF – tem melhor aceitação, devido ao aprimoramento da performance e satisfação dos utilizadores (White, 2005).

- **Anotações Semânticas**

Pode-se dizer que Anotação Semântica é um mecanismo que auxilia o computador a entender o significado de um conjunto de dados. Este processo é mais utilizado na área de web semântica, iniciativa da W3C, objetivando adicionar conteúdo ou metadados às páginas de internet afim de promover integração, automação e reuso de dados entre várias aplicações (Agosti *et al.*, 2007). Para atender estes requisitos, as anotações devem ser baseadas em um domínio formal, como uma ontologia, por exemplo.

As anotações semânticas proporcionam a ligação entre os dados

contidos em um documento e a ontologia. Normalmente é uma referência (indexação) para um ou mais termos definidos dentro da ontologia. Existem diversos estudos de métodos para automatizar o processo de anotações semânticas, como por exemplo as iniciativas de (Kiryakov *et al.*, 2004; Reeve e Han, 2005; Ding *et al.*, 2006) devido ao fato de que este é um processo oneroso.

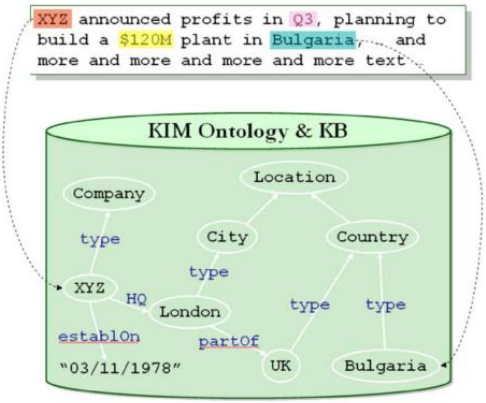


Figura 20 - Ilustração do funcionamento do processo de anotações semânticas auxiliado por uma ontologia Fonte: (Kiryakov *et al.*, 2004)

O processo de anotação semântica se dá pela extração dos nomes ou atributos das entidades em textos não estruturados, como por exemplo, nomes, cargos, entre outros. Após este procedimento, é necessário fazer o relacionamento entre os elementos extraídos do processo anterior com os descritores semânticos introduzidos na ontologia relacionada, para que estas referências possam ser utilizadas para a nomeação das entidades no texto não estruturado.

Dá-se o nome de Repositório Semântico à informação de múltiplos índices de diferentes fontes e vários formatos, e este repositório permite ao usuário efetuar diferentes tipos de busca auxiliados por seu conteúdo.

3.5.8. Ferramentas para Engenharia do Conhecimento

Nesta subseção serão revistas e discutidas ferramentas que oferecem contribuição direta para a realização dos experimentos relacionados mais adiante. Serão brevemente demonstrados os modelos RDF, que neste contexto têm a função de estruturar e armazenar o conhecimento, seguido das ferramentas JENA e Protègè API, que são intimamente

ligadas e têm a funcionalidade de manipular a estrutura e conhecimento contidos em um modelo RDF e, finalmente, as ferramentas da Apache: Lucene e Solr, que trabalham com a indexação e recuperação de informação.

- **RDF**

RDF (*Resource Description Framework*) trata-se de um padrão modelado para prover, através da linguagem XML, uma forma de descrever elementos através de metadados. Foi proposto pela W3C no intuito de descrever de melhor forma dados na web. Para a utilização do RDF para descrever recursos, a utilização de metadados pareceu ideal em conjunto com o modelo RDF a fim de apresentar os recursos nos mais diversos padrões de tipos de comunidades da web mantendo o seu padrão semântico (Lassila e Swick; Dias *et al.*, 2004).

Tem como característica principal a forma com que são descritos os recursos através dos metadados, cujo modelo possui basicamente três tipos de objetos (Dias *et al.*, 2004):

- Recursos: trata-se de todo elemento descrito em RDF, que podem ser objetos em uma página web, uma página web ou até um conjunto de páginas web, definidos por uma URI (*Uniform Resource Identifier*);
- Propriedades: são atributos descritos através de um determinado aspecto utilizado para descrever um recurso;
- Declarações: são declarações, propriamente ditas, envolvendo um nome, uma propriedade e um valor atribuído a ela. Uma declaração é composta por *sujeito*, *predicado* e *objeto*.

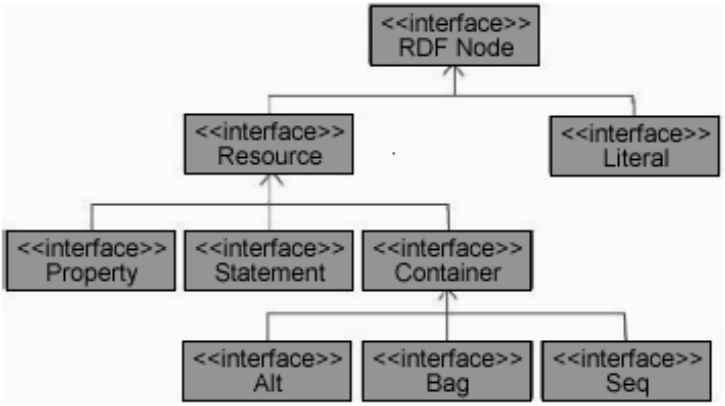


Figura 21 - Ilustração de um modelo RDF
Fonte: (Dias *et al.*, 2004)

A Figura 21 mostra a representação de um modelo RDF, como pode ser visto acima.

Para manipulação dos recursos disponibilizados através da representação de modelos (ou grafos) RDF são normalmente efetuadas por *frameworks* específicos. O mais utilizado destes é nomeado JENA, e será descrito na sequência.

- **JENA**

Jena é um framework desenvolvido na a linguagem JAVA pela Hewlett-Packard Company, por Brian McBride, para dar suporte a aplicações de web semântica para softwares capazes de utilizá-lo (Mcbride, 2002; Carroll *et al.*, 2004; Dias *et al.*, 2004).

Possui uma API que possibilita aos desenvolvedores manipular modelos baseados em RDF. Permite a criação e manipulação de modelos RDF como conjuntos de triplas - (sujeito, predicado, objeto - (Mcbride, 2002; Carroll *et al.*, 2004; Dias *et al.*, 2004).

A sua arquitetura é ilustrada na Figura 22, e demonstra que o coração da aplicação está na API RDF, que é quem suporta a criação, manipulação e consulta em grafos RDF. Seu idealizador (Mcbride, 2002) define a API RDF do JENA conforme transcrito abaixo:

> "Its heart is the RDF API, which supports the creation, manipulation, and query of RDF graphs. The API also supports several different storage technologies. Plug-in interfaces accommodate automatic readers and writers for different languages that developers can use to represent RDF graphs. Above the RDF API sits an inference

layer, query function, and network API" (Mcbride, 2002).

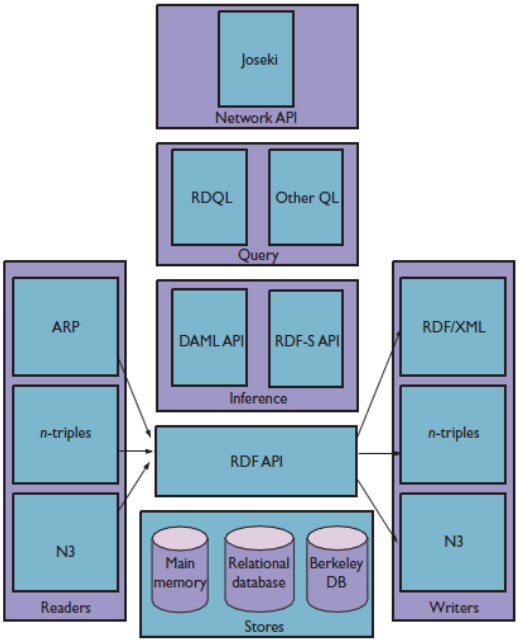


Figura 22 - Arquitetura do Jena. A API RDF é o coração da arquitetura, onde é suportada a criação, manipulação e consulta nos grafos RDF. Fonte: (Mcbride, 2002)

Foi utilizado como framework para o aplicativo de gerenciamento e criação de ontologias, chamado Protègè, que será descrito a seguir.

- **Protègè API**

O Protègè é um ambiente de conhecimento e desenvolvimento de sistemas que vem evoluindo há mais de uma década. Teve sua origem como uma pequena aplicação concebida para atender o domínio médico (protocolo baseado em planejamento terapêutico), mas evoluiu para um objetivo muito mais geral, o de construir um ambiente extensível que atendesse as necessidades mais comuns aos desenvolvedores de aplicações para Web semântica (Gennari *et al.*, 2003; Knublauch, Holger *et al.*, 2004). Mais recentemente, desenvolveu uma comunidade mundial de usuários que têm a possibilidade de adicionar funcionalidades à ferramenta e desta forma contribui para a sua evolução.

Também fornece uma API em JAVA para possibilitar o acesso

e manipulação de ontologias OWL. Esta API basicamente encapsula o mapeamento do conteúdo de uma ontologia em owl. O Plugin OWL fornece um mapeamento detalhado entre sua API estendida e a biblioteca padrão, Jena. Depois de uma ontologia foi carregada em um modelo de Jena, o *Plugin* OWL gera os objetos correspondentes e possibilita sua manipulação (Knublauch, H *et al.*, 2004).

- **Lucene**

Trata-se de uma ferramenta multifuncional de alto desempenho desenvolvida pela Apache (Apache) para facilitar o desenvolvimento de aplicativos que requerem busca *full-text*, podendo ser utilizada em ambientes multi-plataforma.

Sua popularidade se dá pela simplicidade de sua utilização. A exposição das funcionalidades de sua API é um sinal de um software bem desenvolvido. Conseqüentemente, não é necessário ter um conhecimento profundo sobre o funcionamento interno do Lucene para que se possa iniciar seu uso (Hatcher e Gospodnetic).

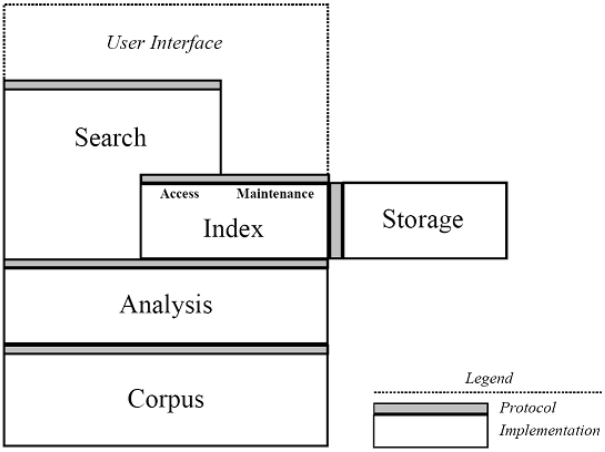


Figura 23 - Arquitetura do Lucene
Fonte: (Gospodnetic e Hatcher, 2004)

- **Solr**

Trata-se de uma plataforma de busca popular e de código aberto do projeto Apache Lucene. Possibilita busca em texto livre, *highlighting*, pesquisa facetada, agrupamento dinâmico, integração com bancos de dados e integração com documentos em formato texto (Solr, 2007), como documentos no formato .doc e .pdf, por exemplo.

Segundo seus desenvolvedores, o Solr é uma ferramenta escalável, fornecendo métodos de busca distribuída e replicação de

índices, recursos de facilidades de navegação, entre outras vantagens.

Sua implementação é feita em Java e roda como um servidor de busca *standalone* dentro de um *servlet*, como o Tomcat, por exemplo.

Utiliza como núcleo de aplicação o já citado Apache Lucene para indexação e recuperação de documentos, e formata como valores de saída retornos em XML ou JSON para facilitar a manipulação em aplicações desenvolvidas que fazer uso deste retorno.

Possibilita configuração externa através de arquivos XML, e possui uma grande gama de *plug-ins* que podem ser acoplados a esta aplicação para incorporar funcionalidades ao Solr.

3.6. TRABALHOS CORRELATOS

Nos dias de hoje, os estudos relacionados aos métodos computacionais de busca textual tornaram-se especialmente voltados ao aperfeiçoamento da web semântica e técnicas para a sua construção. Isto ocorre pelo fato da difusão da web semântica como forma de organizar e recuperar de forma ágil documentos na web, utilizando técnicas computacionais específicas.

Embora não seja o foco da pesquisa, muitas das técnicas da web semântica merecem ser analisadas no sentido de incrementar o arcabouço metodológico das pesquisas que envolvem recuperação de conhecimento.

A seguir estão descritos alguns estudos relevantes sobre recuperação do conhecimento e web semântica, que serão utilizados como embasamento para o desenvolvimento desta pesquisa.

3.6.1. Web semântica

Conforme citado anteriormente, estudos sobre web semântica vêm sendo amplamente discutidos com a intenção de se encontrar a melhor abordagem para o desenvolvimento de aplicações de recuperação do conhecimento na web.

Tim Berners-Lee, idealizador da web semântica, propôs a arquitetura básica que contribui para a definição de técnicas que permitem aos computadores emularem a compreensão do conteúdo da web (Berners-Lee *et al.*, 2001) através da composição de *tags* de metadados, ferramentas de contextualização através destes metadados, entre outros artefatos que permitem que esta afirmação seja verídica.

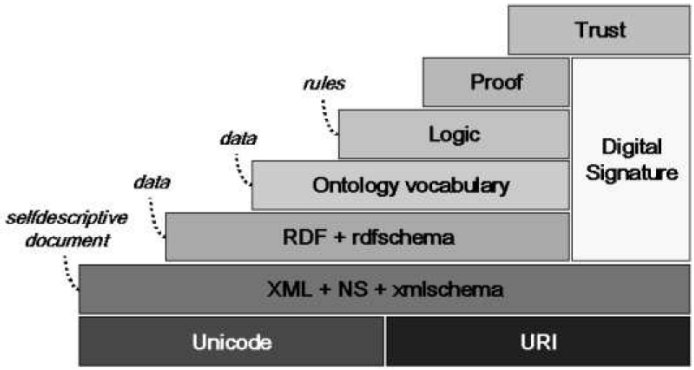


Figura 24 - Modelo de camadas da web semântica
Fonte: (Koivunen e Miller, 2001)

Dentre todas estas pesquisas, existem alguns aspectos básicos que são mantidos por já terem sido previamente classificados desta forma, ou seja, independentemente de sua implementação, componentes como ontologias, *crawlers*, entre outros, já são firmados como sendo indispensáveis para o desenvolvimento de aplicações para a utilização da web semântica.

O que difere entre as abordagens são os métodos para a interoperabilidade entre estes componentes e a forma com que se recupera, de fato, o conhecimento.

Pesquisas abrangentes propõem plataformas híbridas relacionando conceitos de abordagens simbólicas e subsimbólicas, por exemplo. Rocha (Rocha *et al.*, 2004) relaciona estas abordagens em seu estudo "*A Hybrid Approach for Searching in the Semantic Web*", utilizando resultados de pesquisas clássicas de dados associados a conceitos da ontologia. Este resultado passa a ser entrada de dados para a utilização da "ativação por *spreading*" - onde se encontra a abordagem híbrida - em que, a partir da ligação com algumas características de uma dada ontologia, relacionam e definem pesos para os relacionamentos constantes na própria ontologia. Neste caso, a proposta de diferenciação dos modelos clássicos ficam mais evidentes na implementação do algoritmo de "ativação por *spreading*".

Wang (Wang e Jhuo, 2009), por sua vez, procura resolver problemas comuns às implementações da busca Facetada. Busca facetada é um assunto bastante comum na área de Ciências da Informação, e segundo Duarte (2010) trata-se de "um tipo de classificação capaz de identificar características comuns a diversas categorias de um assunto, organizando-o em partes denominadas de

facetas", e serve para decompor o assunto em subclasses até que as variações se esgotem, e após este processo concluído, há o mapeamento destas facetas por áreas iniciando desta forma a composição de estruturas semânticas. Em seu estudo sobre busca facetada no contexto da websemântica, Wang foca na integração de ontologias disponíveis na web de forma a garantir integridade em um determinado domínio. Ainda segundo Wang, com a implementação da camada de busca facetada é possível acentuar a classificação e organização das ontologias.

Ilyas (Ilyas *et al.*) propõe um modelo conceitual de um mecanismo de recuperação utilizando web semântica composto por mecanismos de inferência, *crawlers*, ontologias de mapeamento, entre outros. Porém, nesta proposta os autores procuram trabalhar com maior afinco no mecanismo de inferência, a fim de tratar algumas lacunas não observadas por outras abordagens. Eles sugerem a implementação do mecanismo de inferência pela linguagem PROLOG justificando que esta seria apropriada devido à grande utilização para raciocínio lógico nos últimos anos. Com isso, os autores afirmam terem resolvido muitas das lacunas de ordem lógica por eles citadas no estudo.

3.6.2. Busca por linguagem natural

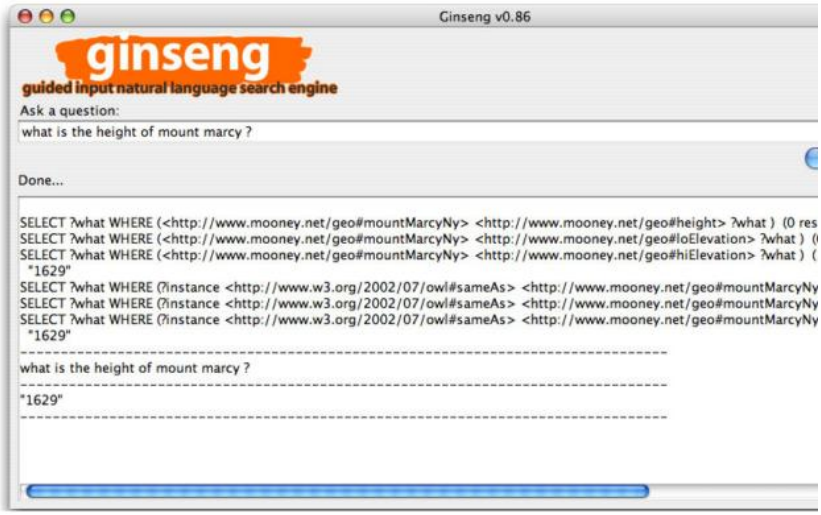


Figura 25 - Interface do usuário do software Ginseng apresentada após uma consulta
Fonte: (Bernstein *et al.*, 2006)

Houve também estudos relacionados à busca por linguagem natural. Um dos estudos mais promissores com relação à linguagem natural utilizando ontologias resultou na implementação do GINSENG

(*A Guided Input Natural Language Search Engine*), por (Bernstein *et al.*, 2006). Trata-se de um sistema que faz buscas por linguagem natural em uma ontologia de domínio, com o diferencial de sugerir a sintática a ser introduzida para a busca com base em um dicionário pré-elaborado.

Desta forma, a consulta seria construída de uma forma compreensível ao depurador de consultas para que o resultado pudesse ser retornado.

A interface de utilização do Ginseng é mostrada na Figura 25.

3.6.3. Busca por Arquivos Invertidos

Além deste estudo relacionado diretamente a ontologias, existem ferramentas capazes de buscar informações em texto livre, normalmente implementadas a partir de técnicas de estruturas de dados conhecidas como "índices invertidos" ou "arquivos invertidos". Uma das mais famosas trata-se da ferramenta Lucene (2005) da Apache, previamente citada, cuja acurácia dos resultados é comprovada em muitos estudos.

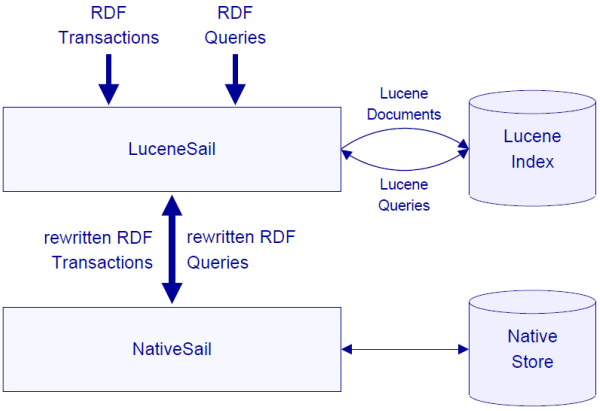


Figura 26 - Ilustração em camadas da arquitetura da extensão em texto livre (Sesame LuceneSail)
Fonte: (Minack *et al.*, 2008)

Partindo deste ponto de vista, alguns autores optaram pela utilização deste tipo de abordagem na recuperação de conhecimento. Tang (Tang *et al.*) utiliza uma ferramenta de código aberto, de produtividade e indexação cuja implementação dá-se sobre o Lucene, chamada Solr (Solr), também da empresa Apache. Em conjunto com esta ferramenta, o autor utiliza o Carrot para incrementar a diversidade de resultados da pesquisa. Segundo Tang, o Carrot colabora para a categorização dos elementos a partir de várias opções de algoritmos. O produto final desta integração é utilizado para a categorização e

indexação para recuperação de informações sobre imagens, cuja implementação foi chamada pelos autores de ImageCLEFPhoto.

Outro exemplo de integração de ferramentas open-source é demonstrado nas pesquisas de Minack (2008) onde também é feita a utilização do Lucene em integração com outras ferramentas. O objetivo deste trabalho é possibilitar a busca estruturada e em texto livre. Para isso, o autor lança mão também da ferramenta Sesame (Broekstra *et al.*, 2002), que é utilizada para fazer a manipulação dos padrões RDF. A partir desta integração, podem ser convertidos os padrões de um documento RDF no formato de documentos do Lucene (ilustrada na Figura 34), possibilitando sua indexação e posterior recuperação. Segundo os autores, o tempo de resposta com esta indexação dos grafos melhorou consideravelmente, embora tenham relacionado como trabalhos futuros a resolução de algumas limitações do sistema devido ao design e abordagem dos sistemas envolvidos serem diferentes. A estrutura da integração das ferramentas utilizadas no trabalho de Minack pode servir como base para a aplicação da integração das ferramentas a serem utilizadas neste trabalho.

Alguns estudos também identificam a possibilidade de apoiar-se na estrutura de alguns Sistemas Gerenciadores de Bancos de Dados - SGBD's. Heese (Heese *et al.*, 2007) afirma que, mesmo algumas tecnologias já tendo a possibilidade de indexar diretamente padrões de padrões RDF, porém esta indexação resulta em alguns casos na perda da informação estrutural contida nelas.

3.6.4. Expansão de consulta

Como previamente descrito no estado da arte, o objetivo da expansão de consulta é o aperfeiçoamento do modelo tradicional de Recuperação de Informação, no sentido de ampliar o universo de busca através da composição de consultas mais completas e abrangentes.

Os pesquisadores Díaz-Galiano e Martín-Valdivia (2009) propuseram uma abordagem onde se utiliza como base de conhecimento adicional para a composição de consultas através da navegação no *Medical Subject Headings* (MeSH) (Nlm, 2006), enquanto Yoo (Yoo e Choi, 2010) utiliza como base de conhecimento adicional o MEDLINE para expandir consultas. Estas abordagens foram desenvolvidas para auxílio nas pesquisas na área da saúde, portanto, tornam-se relevantes para se tomar como parâmetro no desenvolvimento de uma solução para a área de Toxicologia para a língua portuguesa. Uma ilustração do fluxo

utilizado pelo mecanismo de composição de consultas utilizando uma base externa pode ser visualizada na Figura 27.

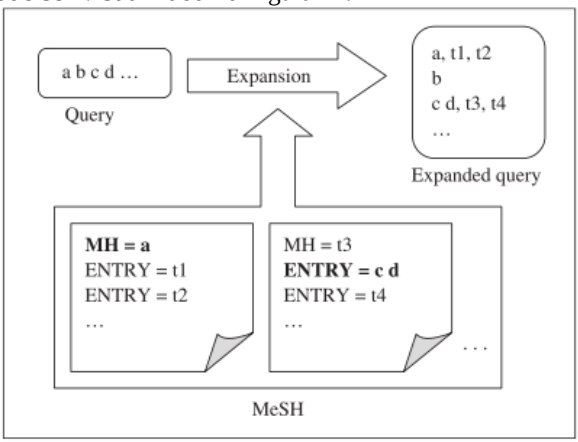


Figura 27 - Expansão de Consulta utilizando os termos do MeSH
Fonte: (Díaz-Galiano *et al*., 2009)

Uma outra abordagem sobre expansão de consultas é a de Chang (2007), que se trata de um mecanismo de Expansão de Consulta mais genérico, que utiliza regras *fuzzy* para inferir quais são os graus de relevância dos termos da consulta digitada pelo usuário com auxílio da freqüência de termos relevantes e a freqüência inversa de um documento (idf) de um termo. Embora seja bem embasada e produza bom resultados num contexto genérico, não se adapta perfeitamente ao contexto aplicado em nosso estudo, servindo como parâmetro adicional no desenvolvimento da ferramenta de expansão de consulta.

3.7. CONSIDERAÇÕES FINAIS

Neste capítulo contextualizou-se, através da revisão de literatura, o objetivo do trabalho em relação à Engenharia do Conhecimento, tendo em vista a utilização de técnicas desta disciplina para apresentar soluções para o problema de pesquisa.

Na inicialização do capítulo, foi possível observar como vêm sendo utilizada a Engenharia do Conhecimento na área da saúde, sua aplicação e benefícios trazidos através de seus métodos e técnicas, ilustrados no emprego dos Softwares de Apoio à Decisão Clínica.

Após isto, foram vistos os aspectos do problema de heterogeneidade e dispersão do conhecimento de mesma natureza em um âmbito genérico. Exemplificações e estudos de caso foram

demonstrados através de citações de autores sobre a importância da integração e comunicação entre bases de conhecimento com características similares, possibilitando a utilização coletiva de modo a poder criar, manipular e disseminar o conhecimento.

Subsequentemente foi apresentado o domínio no qual se enquadra o estudo de caso a ser realizado para verificar a viabilidade da construção da solução proposta para a pergunta de pesquisa. Foram apresentados o ambiente do Centro de Informações Toxicológicas, fluxo de atendimento prestado por um profissional atuante na área, encaminhamento de fichas de caso e pesquisa por agentes intoxicantes envolvidos em um caso.

A partir desta contextualização, foi possível confirmar a necessidade da aplicação de Engenharia do Conhecimento no processo de atendimento urgência prestado no CIT, para facilitar não só a execução das atividades intensivas em conhecimento realizadas pelos plantonistas, como também para possibilitar que este atendimento fosse feito de maneira a oferecer risco menor ao paciente devido à agilidade no encontro do conhecimento necessário ao diagnóstico.

Dispostas estas necessidades, foram então apresentados os métodos e técnicas de Engenharia do Conhecimento pertinentes a uma possível estruturação de solução para o problema de pesquisa, baseado na utilização destes recursos para resolução problemas semelhantes.

Na categoria de ferramentas foram apresentados elementos para a constituição e organização estrutural para armazenamento do conhecimento, como as estruturas de modelos RDF, ferramentas de manipulação destes modelos (JENA), interface (framework) de manipulação e integração do conhecimento (API do Protègè). Também foram apresentadas técnicas de Recuperação de Informação, intrínsecas nas ferramentas que contém as funcionalidades proporcionadas por estas técnicas são Lucene e Apache Solr, uma contida em outra respectivamente.

Na categoria de métodos, além da breve contextualização e demonstração das possíveis metodologias a serem utilizadas para construção de um **Sistema de Conhecimento**, como expressão das atribuições inerentes a uma metodologia foi utilizada para exemplificação a metodologia **CommonKADS**, apresentando resumidamente fluxos, parâmetros e atributos apresentados na sua definição e aplicação.

Técnicas mais elaboradas através da composição de ferramentas foram apresentadas na sequência através da exposição da utilização de

Recuperação da Informação, técnicas de Expansão de Consulta e utilização de Anotações Semânticas.

Na seção **Trabalhos Correlatos** foram apresentados estudos de aplicação relacionados aos métodos e técnicas previamente demonstrados, classificados nas áreas de web semântica, busca por linguagem natural, busca por arquivos invertidos e expansão de consluta, que possuem algum argumento ou artefato necessário à implementação da estrutura da proposta que será descrito na seção seguinte.

Na Tabela 8 são tabulados os trabalhos relacionados levando em consideração os requisitos que se deseja atacar na proposta que será apresentada mais adiante. A Tabela 7 descreve o significado das legendas utilizadas para identificação das funcionalidades.

Tabela 7 - Descrição das legendas dos trabalhos relacionados

Legenda	Descrição
IDX	Abordagens que utilizam alguma técnica de indexação.
SM	Abordagens que utilizam técnicas de implementação semântica.
PLN	Abordagens que utilizam buscas por linguagem natural.
QE	Abordagens que utilizam expansão de consulta.
IBDH	Utilização de integração de bases de dados heterogêneas.

Tabela 8 - Tabulação das funcionalidades/trabalhos relacionados

Autor	IDX	SM	QE	IBDH
(Berners-Lee *et al.*, 2001)		x		x
(Rocha *et al.*, 2004)		x		
(Wang e Jhuo, 2009)		x		
(Ilyas *et al.*)		x		x
(Bernstein *et al.*, 2006)				
(Tang *et al.*)	x			
(Minack *et al.*, 2008)	x			
(Heese *et al.*, 2007)	x			
(Díaz-Galiano *et al.*, 2009)	x		x	
(2007)			x	
(Yoo e Choi, 2010)			x	

4. PROPOSTA

Neste capítulo serão apresentados e discutidos os itens relacionados à proposta de resolução do problema proposto neste estudo. Serão apresentados conceitualização, funcionalidades e desenvolvimento de cada um dos módulos implementados no Sistema de Conhecimento proposto. Ao final são apresentadas as considerações finais sobre a estruturação e construção do protótipo deste Sistema de Conhecimento.

4.1. CONCEITUALIZAÇÃO

O primeiro aspecto importante a ser levado em consideração e identificado no modelo de **Contexto Organizacional** do **CommonKADS** (Schreiber, 2000) é o fato de que a proposta deste estudo é desenvolver um Sistema de Conhecimento de **Apoio à Decisão Clínica**, e **não** desenvolver um sistema especialista que tenha a capacidade de tomar decisões. Muitos são os fatores para se tomar a precaução de não desenvolver um sistema com este tipo de ambição, entre eles a dificuldade de se fazer uma anamnese correta. Kong cita alguns destes fatores:

> "Uncertainty exists in almost every stage of a clinical decision making process. Sources of uncertainties may include that patients can not describe exactly what has happened to them or how they feel, doctors and nurses can not tell exactly what they observe, laboratories report results may be with some degrees of error, physiologists do not precisely understand how the human body works, medical researchers can not precisely characterize how diseases alter the normal functioning of the body, pharmacologists do not fully understand the mechanisms accounting for the effectiveness of drugs, and no one can precisely determine one's prognosis" (Kong *et al.*, 2008).

Esclarecido este ponto importante, iniciou-se o desenvolvimento estrutural da proposta conforme descrito na sequência.

Conforme previamente descrito, o objetivo deste trabalho é resolver o problema da dispersão e heterogeneidade da informação utilizada cotidianamente pelos profissionais dos CIT's, para isso, propôs-se um modelo de integração com funções modulares bem definidas, de forma que possam realizar todas as etapas propostas.

4.1.1. Adesão ao CommonKADS

A aplicação da metodologia CommonKADS neste estudo deu-se pela utilização conceitual dos modelos de Contexto Organizacional – onde foram feitas análises prévias à realização deste estudo por um profissional na área de Engenharia de Software, e foram reutilizadas como parâmetro para a estruturação do trabalho - e Modelo de Tarefa, além de utilizar o conceito de MVC (*Model*, *View*, *Controller*) na arquitetura de construção do sistema, separando a camada de conhecimento das camadas de interface e aplicação.

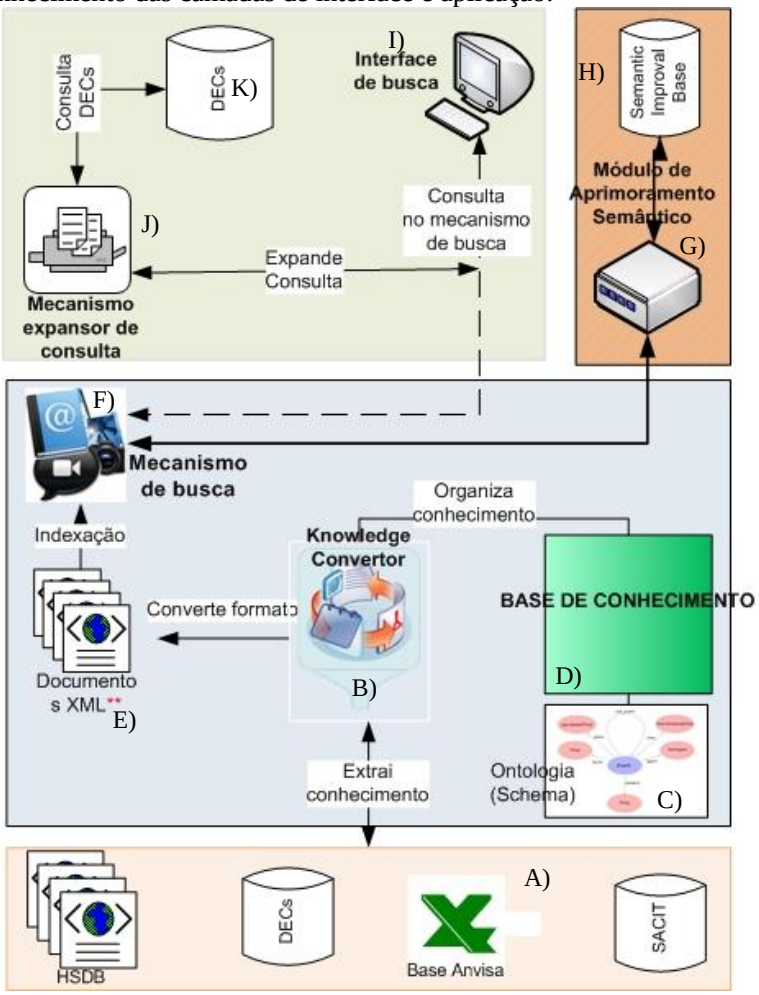


Figura 28 - Ilustração da estrutura do mecanismo de busca proposto

A fragmentação esboçada na Figura 28, através de diferentes cores e letras, retrata a modularização da proposta. Para melhor visualização, a distinção entre processos e artefatos é feita através do formato dos indicadores, através de letras. Eis, detalhadamente, a descrição destes fragmentos:

A) Bases de dados: são artefatos que contém grande parte a informação utilizada cotidianamente pelos profissionais que prestam atendimento no CIT. Cada uma das bases descritas possui dados referentes a um determinado contexto;

B) *Knowledge Convertor* – função de Mecanismo de integração: trata-se de um mecanismo desenvolvido para converter o formato de conhecimento contido em um tipo de base de dados para outro. Em sua utilização inicial, este mecanismo é responsável pela integração das bases de dados citadas anteriormente. Cada uma destas bases possui particularidades e uma estrutura planejada para um fim específico. Desta forma, existe a necessidade de estabelecer um padrão para que todas as bases de dados possam ser empregadas na construção de uma única base de conhecimento. Para que isso seja possível, este padrão deve ser definido através de uma ontologia (C), que será descrita a seguir. A partir de então, inicia-se o processo da integração, elucidado na sequência:

    1. Uma vez definido o padrão, o mecanismo de integração deve ter a capacidade de ler os dados de cada uma destas bases, e converter para o padrão previamente definido. Esta leitura é feita através de *crawlers* específicos desenvolvidos especialmente para esta função. Os *crawlers* acessam a base de dados definida e fazem a extração, armazenando em uma base de dados local esta informação.
    2. O mecanismo de integração, guiado pela ontologia (C) faz a exportação dos dados para uma base de conhecimento (D), seguindo as definições previamente definidas.

C) Ontologia (Schema): conduz o processo de exportação dos dados para a base de conhecimento, de forma que esta possua um formato homogêneo para instanciação das entidades que até então estavam dispostas em formato heterogêneo e sem conexão entre si;

D) Base de conhecimento: artefato produzido como resultado da integração das bases de dados (A), com auxílio do mecanismo de integração (B). Possui uma classificação definida pela ontologia (C), onde seus indivíduos ou instâncias estão organizados de maneira hierárquica. Estes indivíduos possuem entre outras, informações diversas sobre substâncias, sintomas, princípios ativos de composição, além de relacionamentos entre si para que o universo de conhecimento produzido reflita a realidade do domínio;

B) (2) *Knowledge Convertor* – função *Parser* de conversão: nesta função, o mecanismo de conversão é responsável pela conexão entre a base de conhecimento (D) e o mecanismo de busca (F). Este processo é descrito da seguinte forma:

> 3. O módulo faz a conexão com a base de conhecimento, extrai este conhecimento;
> 4. O conhecimento é convertido em um padrão legível ao mecanismo de busca, sem que seja afetada a semântica das ligações previamente dispostas no domínio;

E) Documentos XML: são gerados pelo referido módulo de conversão (*Knowledge Convertor*). Possuem uma estrutura pré-definida pelo indexador do mecanismo de busca, e possuem basicamente a mesma estrutura semântica definida no domínio da base de conhecimento (D);

F) Mecanismo de busca: é este o mecanismo responsável pela indexação (5) e pesquisa (6) dos dados da base de conhecimento (D). Nesta pesquisa, foi utilizado como mecanismo de busca um aplicativo open source amplamente utilizado para estes fins, denominado Apache Solr (Solr). Os processos pelos quais este mecanismo é responsável são:
5. indexação: o Solr faz a leitura dos documentos XML (E) que estão dispostos em uma estrutura específica, padronizada pela Apache. Este processo é todo efetuado utilizando a implementação padrão do Solr;
6. recuperação: processo que tem seu gatilho na interface de busca, no momento em que o usuário aciona esta opção. Este processo é guiado, além da interface e de um controlador específico para esta questão, por um módulo de

expansão de busca (J). Desta forma, através de uma consulta simples com palavras-chave ou orações completas, uma consulta expandida é feita no mecanismo de recuperação, o qual retorna os resultados baseado nos algoritmos disponíveis na implementação do Solr, com o auxílio do

G) Módulo de Aprimoramento Semântico: que se trata de um *plug-in* adaptado ao Solr com o intuito de estabelecer relacionamentos entre os conceitos definidos por um grau de similaridade, definido no momento da indexação. Este módulo utiliza uma base de conhecimento auxiliar

H) que contém a informação sobre a similaridade entre os conceitos, que posteriormente auxilia na recuperação de resultados não relacionados nos itens relevantes de uma busca puramente sintática. Têm também como adereço a implementação de um algoritmo de fonética em língua portuguesa, também adaptado diretamente ao Apache Solr;

I) Interface de busca: trata-se da interface web desenvolvida para a interação do usuário com o sistema. Com o intuito de promover a utilização de dispositivos *touch screen*, a interface foi desenvolvida utilizando este padrão. Esta interface possibilita ao usuário, através do toque na tela ou seleção com mouse, a ativação de assistentes de composição de consulta. Isso permite que o usuário, ao escolher por exemplo a opção substância, tenha à sua disposição a listagem de substâncias referentes a uma pré-consulta. Os resultados desta pré-consulta são atualizados na medida em que a palavra desejada vai sendo composta com a digitação das letras. Após a consulta ter sido composta, o usuário clica em Buscar e o mecanismo controlador do sistema de busca responsabiliza-se por expandir a consulta e trazer seus resultados. Quando os resultados estão dispostos na tela, o usuário pode escolher um deles para verificar a relevância de um determinado resultado e, se decidir, pode escolher itens relacionados para um refinamento de pesquisa, como por exemplo: substâncias relacionadas, sintomas, entre outros;

J) Mecanismo expansor de consulta: possui também um nome que sugere sua funcionalidade. O processo de trabalho deste módulo se dá por:

7. Recebe como entrada a consulta digitada pelo usuário na interface (I) e a expande através da,

8. consulta sinônimos e itens que se encontram no mesmo grau da hierarquia DeCS (K), concatena estes sinônimos e descritores de mesmo nível juntamente com a consulta estabelecendo prioridade sobre a consulta original o mecanismo de busca (F);

K) Base de dados do DeCS: base de dados hierárquica, com informações sobre diversas áreas da saúde, amplamente utilizada para indexação de assuntos relacionados a esta área. É utilizada como base adicional para expansão da busca baseada na consulta digitada pelo usuário.

4.2. FUNCIONAMENTO

No modelo proposto, é possível efetuar uma busca em um local centralizado e obter grande quantidade das informações necessárias para um atendimento. Isso é possível por meio do desenvolvimento de uma ontologia personalizada com informações toxicológicas que abrangem grande parte deste domínio.

Para isso, basta o plantonista acessar a tela do motor de busca através de uma tecla de atalho. Estarão disponíveis as informações básicas para instruir o usuário sobre como fazer sua pesquisa. Esta tela dispõe de recursos de interface para a utilização de dispositivos sensíveis ao toque, no intuito de facilitar o acesso às funcionalidades, mesmo que o plantonista não disponha das duas mãos livres para a digitação da pesquisa.

Desta forma, no momento em que os dados da consulta são compostos na tela, é possível que o usuário selecione qual ou quais dos campos disponíveis que deseja fazer a busca. A cada filtro selecionado, abre-se uma tela com uma lista de itens e um teclado toutch-screen, para que seja então digitada a palavra a ser buscada. Na medida em que são digitadas as letras, o motor pesquisa por palavras semelhantes ao fragmento digitado, sugerindo desta forma palavras para que o usuário possa selecionar uma destas antes mesmo de digitar a palavra toda, agilizando o processo de composição da consulta.

Após a composição da consulta ser concluída, o motor de busca retorna todas as ocorrências relacionadas a esta consulta por ordem de relevância, estabelecida pelo algoritmo implementado no motor

utilizando as técnicas de expansão de consulta e aprimoramento semântico.
A partir deste produto, cabe ao utilizador do sistema inferir quais são, de fato, os resultados que atendem às suas expectativas, e desta forma continuar o processo de atendimento.

4.3. DESENVOLVIMENTO

Algumas das ferramentas descritas previamente são disponibilizadas em padrão de código aberto para utilização e manutenção. Durante o desenvolvimento desta pesquisa, optou-se pela utilização máxima de recursos já disponíveis, visto que a manutenção destas ferramentas teria maior viabilidade do que o seu desenvolvimento por completo.

Como esperado, algumas especificidades foram encontradas, e por isso fez-se necessário o desenvolvimento de novos módulos para o seu atendimento. Dentre estas especificidades, pode-se citar a necessidade da conversão do conhecimento contido nas bases de dados e posteriormente na base de conhecimento, de forma que pudesse ser indexado pelo Apache Solr.

A integração vista por um aspecto técnico e o desenvolvimento das novas aplicações será mostrado nos tópicos que seguem.

4.3.1. Base de Conhecimento

Após a verificação literária com respeito às metodologias utilizadas para a construção de ontologias, identificou-se que a metodologia apropriada para a construção da base de conhecimento de domínio para Toxicologia Clínica seria a METHONTOLOGY, considerando a definição explícita das fases de criação e reunião de insumos para esta tarefa.

Atendendo aos requisitos da METHONTOLOGY, foram então criados:

- glossário de termos;
- taxonomia de conceitos;
- diagramas de relações entre os conceitos;
- estruturação dos conceitos;
- descrição dos atributos dos conceitos;

- descrição dos axiomas e regras que usam os conceitos estruturados e
- descrição dos indivíduos.

Para isso, foram levantados produtos de informação contidos em vocabulários controlados, tesauros, ontologias e bases de dados referentes à toxicologia clínica e farmacologia com potencial fornecimento de insumos para a base de conhecimento a ser criada.

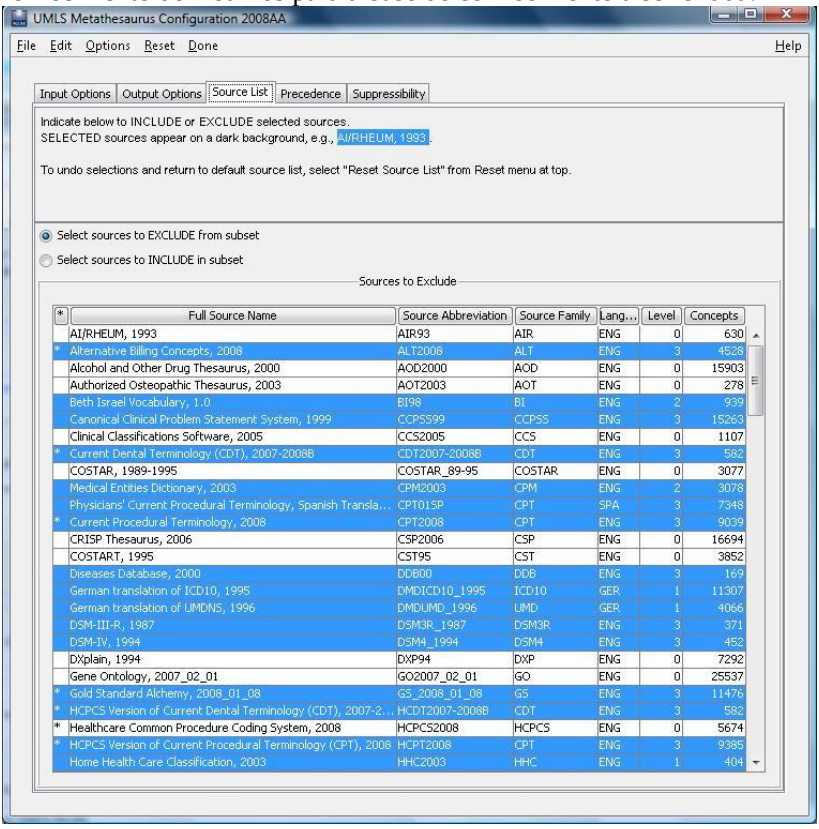


Figura 29 - Seleção de fontes para composição do *subset*

As fontes relacionadas abaixo foram selecionadas para cobrir esta etapa:

- UMLS - Nesta pesquisa, o Metathesaurus foi utilizado para a importação de conceitos e definições pertinentes ao contexto da toxicologia clínica. Foram selecionadas as fontes pertinentes ao projeto e compiladas em um subset, um

subconjunto do Metathesaurus personalizado, e em seguida exportado para um banco de dados.

- A análise do Metathesaurus foi realizada por meio da instalação local do repositório de conceitos e da interface de navegação MetamorphoSys. Durante a instalação do repositório de conceitos, é possível especificar um 'Subset', que é um sub-conjunto da totalidade do repositório de conceitos. Esta especificação é feita por meio da seleção das fontes que se deseja incluir no Subset, selecionando em cada fonte do Metathesaurus (quando disponível) versões ou idiomas específicos.

- O reuso dos vocabulários selecionados pode ser feito no momento da geração do Subset, pois a interface de instalação do UMLS permite exportar um script em SQL, em Oracle ou MySQL, para criar a estrutura de tabelas e índices e popular com os respectivos dados.

- Na Figura 29 é possível visualizar uma das telas do MetamorphoSys.

Nesta etapa da preparação do subset são escolhidas as fontes (vocabulários e tesaurus) que irão compor o conjunto de dados exportados. A partir da UMLS, foi possível compor estruturas de conceitos e suas definições, termos equivalentes e outras estruturas necessárias à composição dos conceitos da ontologia do TELE-CIT.

- Base nativa do CIT-SC - estas bases de dados foram analisadas e os sistemas de classificação normalizados para criar a concepção inicial, expressa na forma de uma Taxonomia de Agentes, implementada diretamente na ontologia. Foi definida uma estratégia para validação dos dados e para a atualização de algumas das classificações (ex: a classificação medicamentos foi substituída pela classificação Anatomical Therapeutic Chemical (ATC), objetivando a padronização das categorias em nível internacional.

- Base ANVISA - os dados desta base foram submetidos ao tratamento para transformação em um banco de dados (importação e normalização) para obtenção da lista de substâncias que seriam posteriormente ligadas à Lista DCB, e

conseqüentemente, vinculadas as respectivas monografias das substâncias, compondo assim um grande banco de informações sobre a toxicologia dos medicamentos constantes naquela lista. Cerca de 90% dos produtos possuíam informações completas e foram empregados no processo de importação;

- HSDB – efetuou-se o download do banco de dados, sendo que os dados estão disponíveis em formato XML, contendo documentos com detalhes sobre substâncias (mais de 140 itens de informação em cerca de 5100 documentos) entre os quais foram selecionados apenas o conjunto de informações pertinentes ao projeto, e então preparados e convertidos para um banco de dados relacional;

- Lista ANVISA-DCB - esta lista foi especialmente útil para relacionar os medicamentos do banco de dados da ANVISA, ligando os princípios ativos ali informados com as monografias de substâncias e outras bases por meio do número CAS. Foi possível também transformar a Lista DCB num dicionário português-inglês dos nomes de substâncias, com apoio da base de substâncias ChemIDPlus da NLM;

- AGROFIT - Foi desenvolvido um web crawler para obtenção dos dados e sua conversão para um banco de dados local, uma vez que as informações são disponibilizadas publicamente por meio de consulta item a item. Estes dados tratados compuseram a parte de Produtos Agropecuários da Base de Conhecimento.

Para a organização, classificação, importação e manutenção dos indivíduos da base de conhecimento utilizou-se a ferramenta Protègè, e a linguagem OWL-DL. Postulou-se a utilização de três plug-ins para a realização da tarefa de importação das informações para o Protègè:

- plug-in UMLS tab;
- plug-in DataMaster tab;
- plug-in Excel Import tab (versão 4).

O plug-in Excel Import mostrou-se mais adequado para a tarefa de importação conforme as atividades desta implementação. Por meio dele foram carregadas e populadas as classes criadas no Protègè, com base na elaboração de visões no banco de dados construídas

especificamente para este fim. A interface deste plug-in pode ser vista na Figura 31.

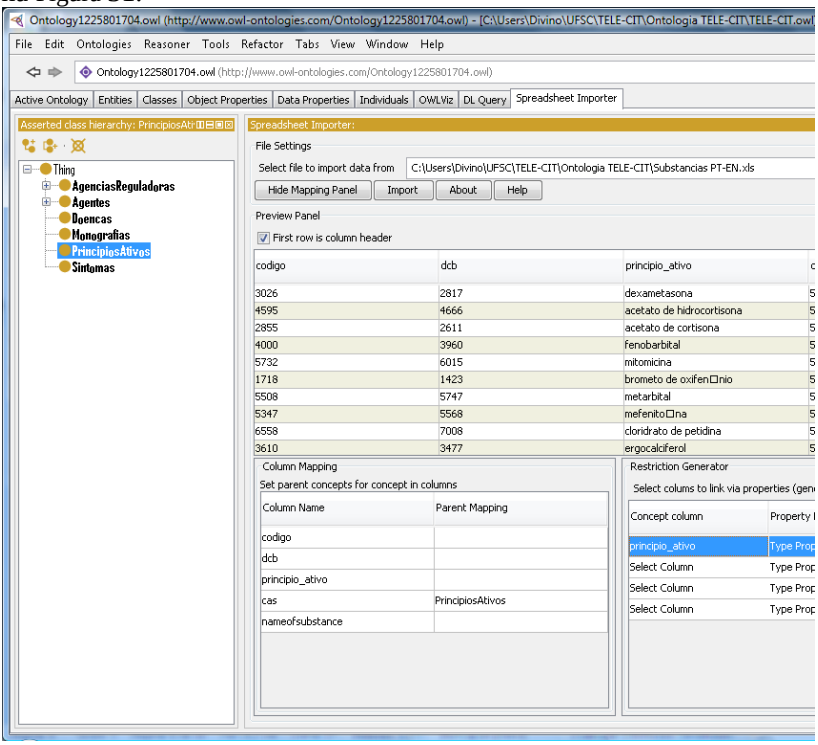


Figura 30 - Excel Import

Como visto anteriormente, a preparação da informação necessária para a construção da base de conhecimento foi realizada por fases, iniciando pela taxonomia de classes, que conteve apenas as definições conceituais dos indivíduos que consistiriam na base de dados. Após isso, a fim de utilizar estrategicamente o recurso para importação de informações para compor a ontologia (criação dos indivíduos da ontologia), foi preparado um conjunto de visões de dados a partir das fontes pesquisadas.

Cada fonte teve seus dados coletados e importados para um banco de dados intermediário, que seria o ponto de partida para a elaboração de visões de dados, sendo que cada uma será importada por meio do plug-in Excel Import do Protege. As fontes foram organizadas em dois bancos de dados que totalizam cerca de 36.400.000 de registros, sendo que o primeiro deles é específico para os recursos da UMLS, e o

segundo contém todas as outras fontes. A Figura 31 contém a ilustração do formato destes bancos de dados.

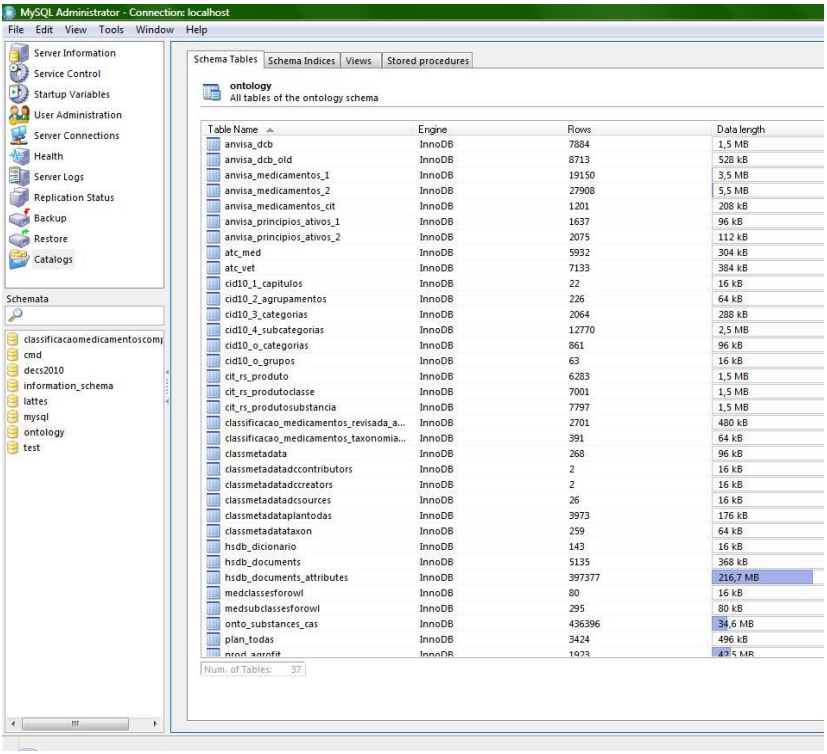


Figura 31 - Banco de Dados com as Tabelas intermediárias para preparar a Ontologia

Por fim, foi desenvolvida uma ferramenta em Java chamada Knowledge Convertor, destinada a popular as classes com os indivíduos previamente preparados a partir das fontes mapeadas e selecionadas. Esta ferramenta será descrita em detalhes mais adiante. Ilustrações do primeiro artefato da base de conhecimento gerado pelo Knowledge Convertor e da classificação final deste artefato podem ser vistas nas Figura 32 e Figura 33.

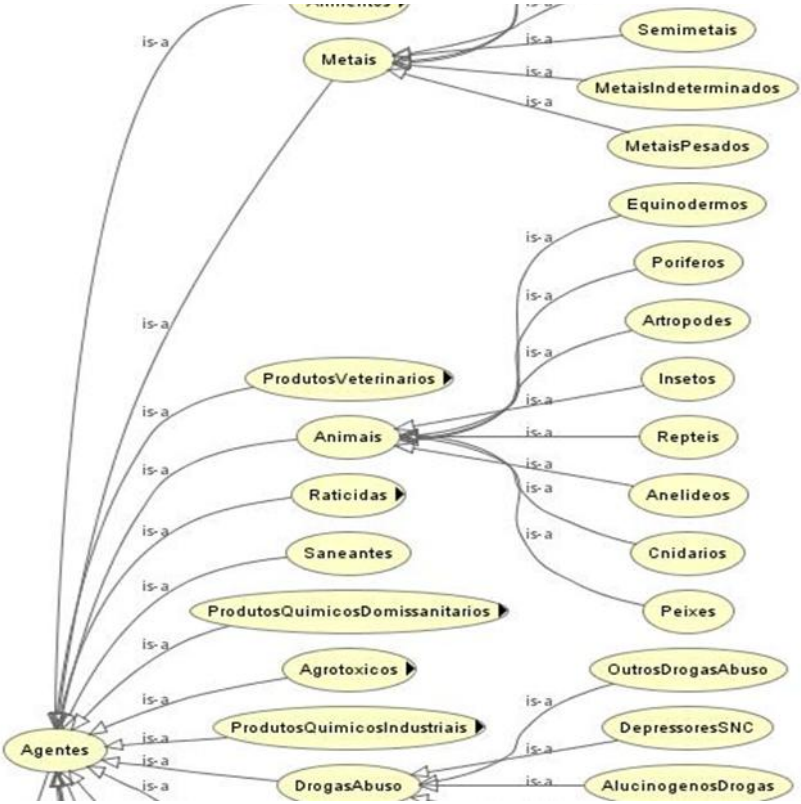


Figura 32 - Visualização de parte da Ontologia do TELE-CIT do TELE-CIT

```
<owl:Class rdf:ID="C009_5309">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="C009"/>
  </rdfs:subClassOf>
  <rdfs:label xml:lang="EN">null</rdfs:label>
  <rdfs:label xml:lang="PT-BR">null</rdfs:label>
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >Preparações de ferro</rdfs:comment>
</owl:Class>
<owl:Class rdf:ID="C076">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="Medicamentos"/>
  </rdfs:subClassOf>
  <rdfs:label xml:lang="EN">null</rdfs:label>
  <rdfs:label xml:lang="PT-BR">null</rdfs:label>
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >Urológicos</rdfs:comment>
</owl:Class>
```

Figura 33 - Tela com fragmento de Codigo OWL da Ontologia

4.3.2. Knowledge Convertor Parser

Assim foi nomeada a ferramenta que é responsável pela conversão do formato do conhecimento contido nas bases de dados distribuídas e heterogêneas que foram selecionadas para a composição da base de conhecimento. Também foi utilizado para a conversão da persistência da base de conhecimento, colaborando para a indexação e armazenamento no Apache Solr, por meio da criação de arquivos XML estruturados.

Seu desenvolvimento deu-se pela utilização da API do Protègè (Knublauch, Holger *et al.*, 2004), implementado sobre o framework JENA (Mcbride, 2002).

Após a construção da Base de Conhecimento com o auxílio do módulo Knowledge Convertor, observou-se a necessidade de uma interface entre a ontologia e o Apache Solr, cujas funcionalidades veremos logo após. Assim sendo, decidiu-se utilizar a ferramenta desenvolvida previamente para atender também a este requisito, visto que o framework utilizado seria o ideal para efetuar a extração do conhecimento da ontologia.

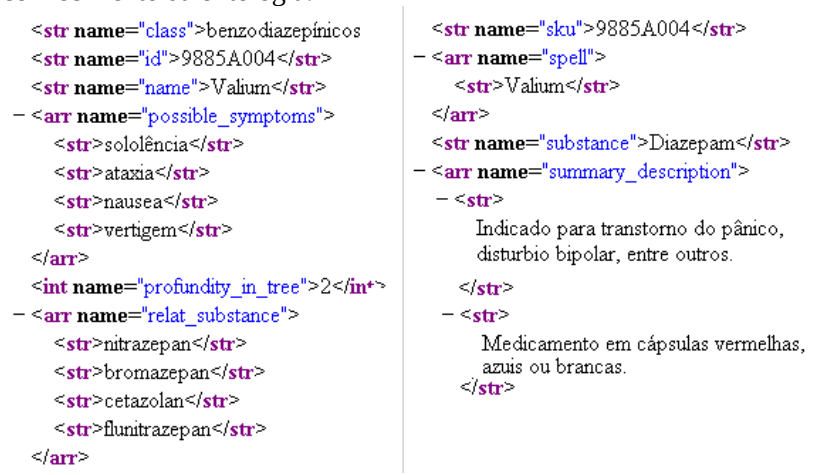


Figura 34 - Fragmento de um documento gerado pela conversão do Knowledge Convertor

Para isso, utilizou-se a mesma estrutura, adaptando somente o algoritmo básico para efetuar a função de extração do conhecimento. No algoritmo foram utilizadas técnicas de recursividade, para que fossem percorridas todas as classes da base de conhecimento, a fim de recuperar todas as instâncias e relacionamentos contidos nas classes. Um exemplo de documento gerado pelo Knowledge Convertor pode ser visto na Figura 34.

Também um fragmento do código desenvolvido para percorrer a ontologia pode ser visto na Figura 35.

```
Statement stm = conn.createStatement();

ResultSet rs1 = stm.executeQuery(SqlSubstanciasToIndividuals);
while (rs1.next()) {
    OWLIndividual newIndividuo = substancias.createOWLIndividual(rs1.getS
    newIndividuo.setPropertyValue(nomePropertyPT, rs1.getString("substanc
    newIndividuo.setPropertyValue(nomePropertyEN, rs1.getString("substanc
    newIndividuo.setPropertyValue(casProperty, rs1.getString("cas"));
    newIndividuo.setPropertyValue(dcbProperty, rs1.getString("dcb"));
    newIndividuo.setPropertyValue(annCodigoOrigem, rs1.getString("codigo"
}

ResultSet rs2 = stm.executeQuery(SqlMedicamentosNivel1);
while (rs2.next()) {
    OWLNamedClass item = onto.createOWLNamedSubclass(rs2.getString("codCl
    item.addComment(rs2.getString("classe"));
    item.addLabel(rs2.getString("PrimeiroDedefinitionPT"), "PT-BR");
    item.addLabel(rs2.getString("PrimeiroDedefinitionEN"), "EN");
}

ResultSet rs3 = stm.executeQuery(SqlMedicamentosNivel2);
while (rs3.next()) {
    OWLNamedClass supitem = onto.getOWLNamedClass(rs3.getString("subClass
    OWLNamedClass item = onto.createOWLNamedSubclass(rs3.getString("codSu
    item.addComment(rs3.getString("subclasse"));
    item.addLabel(rs3.getString("PrimeiroDedefinitionPT"), "PT-BR");
    item.addLabel(rs3.getString("PrimeiroDedefinitionEN"), "EN");
}
```

Figura 35 - Tela com fragmento de Código Java do gerador do Knowledge Convertor

4.3.3. Motor de Busca

O motor de busca utiliza como núcleo a aplicação de código aberto Apache Solr. Esta aplicação permite a indexação e recuperação de itens de forma semi-automatizada. Existem duas formas de fazer a indexação no Solr: a primeira delas, é através de documentos estruturados na linguagem XML, de forma que hajam marcadores de início e fim de documentos, assim como tags que definam quais são os atributos de cada documento.

Um aspecto importante da utilização do Solr neste trabalho, é que ele não foi utilizado unicamente para indexar os documentos, mas também armazenar o seu conteúdo na base de índices, com o intuito de

minimizar a quantidade de operações a serem realizadas no momento da busca.

Desta forma, pode-se dizer que a persistência da ontologia foi comutada do formato owl, para o formato de índices do Solr. Para que a ontologia conservasse este aspecto, optou-se pela utilização de atributos como descritores de relacionamentos, ao invés das propriedades de objeto comumente usadas para esta finalidade.

A partir da indexação das instâncias já se possibilitou fazer a busca direta, porém, uma das necessidades do CIT ainda não era atendida, a busca fonética.

O Solr contém por padrão algoritmos para busca fonética, mas nenhum que se adequasse à Língua Portuguesa. Para isso, desenvolveu-se este algoritmo que foi implantado no núcleo do software, que nativamente permite esta ação.

4.3.4. Módulo de Aprimoramento Semântico

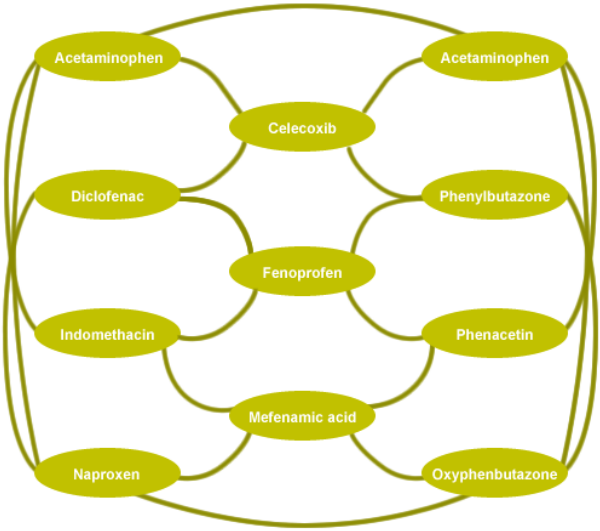


Figura 36 - Ilustração de base de dados auxiliar utilizada pelo módulo de Aperfeiçoamento Semântico

Este módulo foi desenvolvido baseado no plug-in responsável pela sinonímia do Apache Solr. Utiliza uma base de conhecimento auxiliar que possui informação sobre os relacionamentos entre as instâncias, cuja similaridade é definida por um especialista de domínio. Uma ilustração

do modelo de base de dados auxiliar para aprimoramento semântico pode ser vista na Figura 36.

No momento em que as palavras-chave são enviadas ao mecanismo, este trata de fazer a comparação e encontrar os registros relevantes para retorno ao usuário, utilizando como entrada os dados digitados pelo usuário tratados por um algoritmo de fonética desenvolvido para a língua portuguesa.

O conjunto resultante é comparado com a consulta composta gerada pelo mecanismo de expansão de consulta, que será descrito no próximo tópico, a fim de retornar o máximo de resultados relevantes para a busca.

4.3.5. Expansão de consulta

Para incrementar o processo de busca, optou-se por expandir a consulta originalmente realizada pelo usuário. Para isso, a abordagem escolhida foi a de Díaz-Galiano e Martín-Valdivia (2009), que utilizaram esta técnica para expansão de consultas com bons resultados.

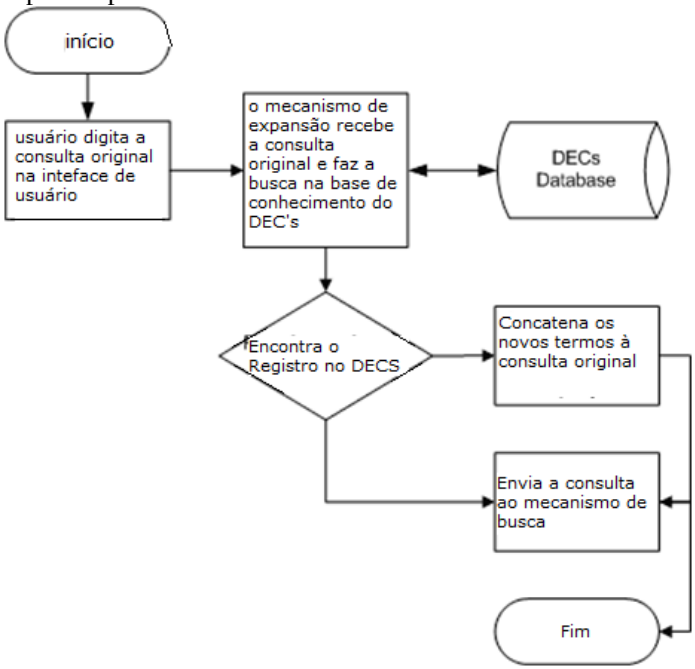


Figura 37 - Fluxo de trabalho do mecanismo de expansão de consulta desenvolvido neste estudo

Porém, para este trabalho, foi preciso adaptar a técnica devido ao fato de que, na abordagem de Galiano, a língua utilizada era a inglesa. Devido a este fato, nesta abordagem utilizou-se como base de dados adicional para expansão de consulta o DECs, por ser uma base trilíngue, contendo descritores em português, espanhol e inglês. Com isso, foi possível pesquisar também na base HSDB, cujos registros encontram-se também na língua inglesa.

Outras adaptações também foram feitas no método em que a árvore das bases adicionais são percorridas. Isto ocorre porque a base DECs possui registros que proporcionam navegação horizontal, ou seja, é possível saber quais são os "irmãos" de um nodo a partir do próprio nodo. Assim sendo, não é necessário subir um nível na árvore para saber quais são seus filhos, para então estabelecer os irmãos do nodo pesquisado.

O fluxo do modelo de expansão de consulta adaptado neste estudo pode ser visualizado na Figura 37.

4.3.6. Interface

Trata-se do módulo que faz a comunicação do motor de busca com o utilizador. Neste protótipo, a interface foi desenvolvida de modo que possa possibilitar ao utilizador efetuar a manipulação do conhecimento intrínseco na aplicação de maneira rápida e descomplicada.

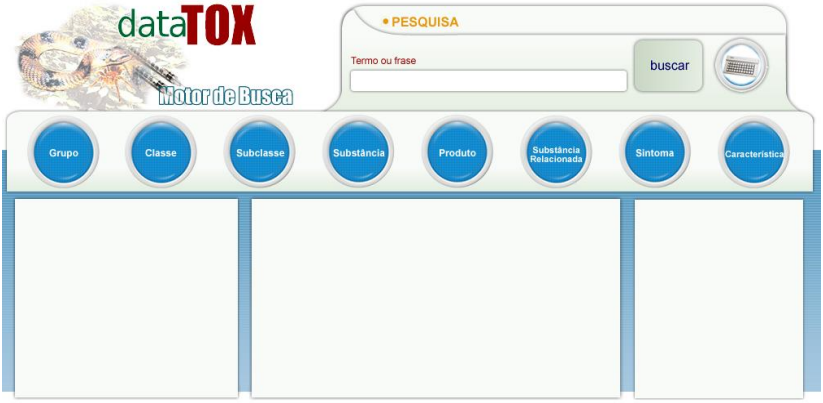


Figura 38 - Interface do protótipo

Levando em consideração que em alguns dos CIT's ainda não utilizam equipamentos com *headset*, objetivou-se implementar uma

interface que pudesse ser manuseada com apenas uma das mãos. Para isso, o *design* da interface foi projetado para possibilitar o uso através de equipamentos com interfaces sensíveis ao toque.

Além desta possibilidade, a interface disponibiliza na tela do usuário opções de navegação por itens similares conforme pré-definido na base de conhecimento, ou seja, se o utilizador tiver interesse, durante o atendimento, na procura por itens semelhantes, ele tem à sua disposição algumas sugestões feitas pelo próprio sistema. A ilustração desta possibilidade de uso pode ser vista na Figura 38.

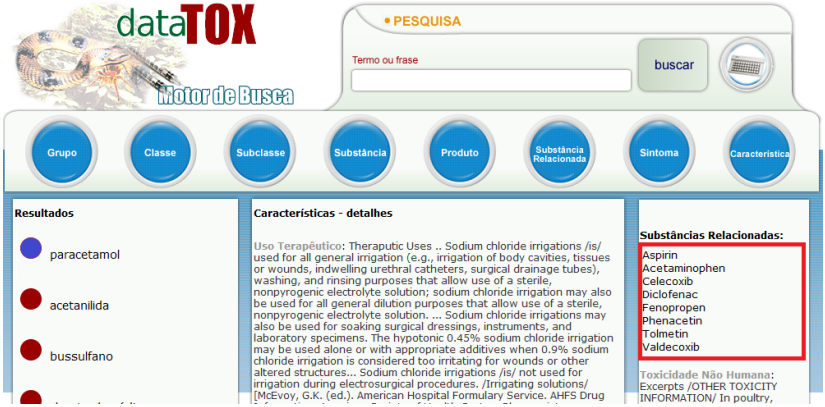


Figura 39 - Ilustração da disposição dos itens relacionados na interface

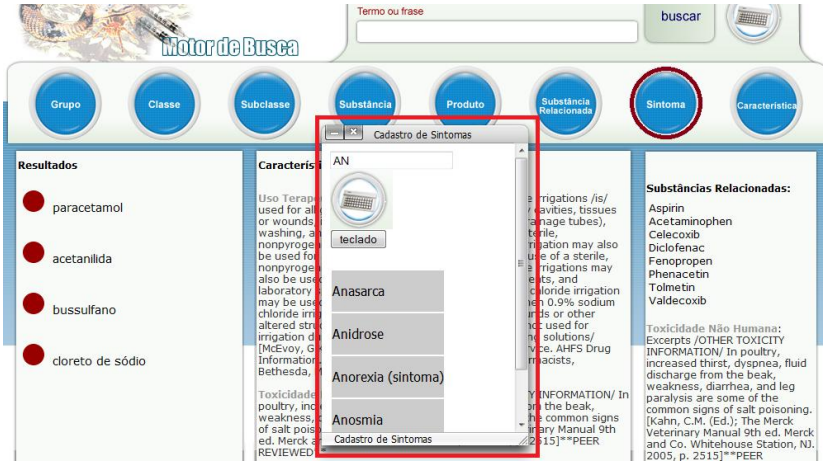
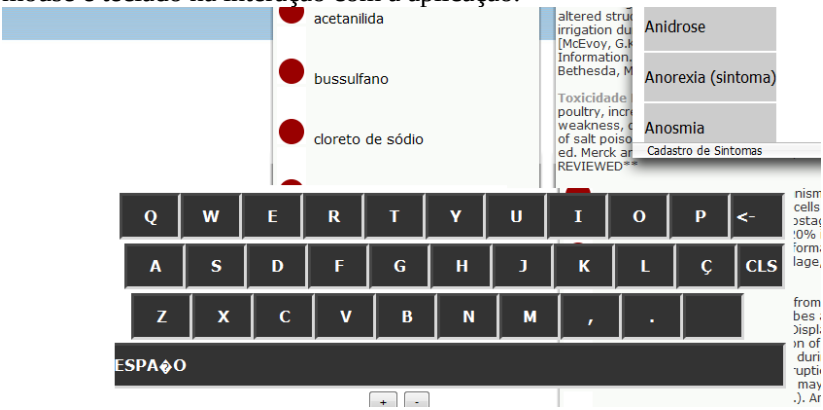


Figura 40 - Demonstração do mecanismo de auxílio à construção de consultas

Também foram desenvolvidos mecanismos auxiliadores para serem utilizados no momento da composição das consultas. Com algumas classificações, podem ser acessados através do teclado virtual, oferecendo sugestões de termos - classes, sintomas, entre outros – para auxiliar na estruturação de uma consulta complexa.

Finalmente, o teclado virtual mencionado anteriormente foi desenvolvido para adequação a dispositivos eletrônicos que não contenham este tipo de adereço no seu software de manipulação. Ele permite a integração com todas as caixas de digitação de texto da interface, possibilitando que não haja a necessidade de utilização de mouse e teclado na interação com a aplicação.



4.4. CONSIDERAÇÕES FINAIS

Nesta seção foram apresentados estrutura e desenvolvimento dos módulos da proposta de motor de busca de conhecimento. A integração dos módulos e o fluxo do conhecimento através deles também foi descrita e exemplificada.

Considera-se que através das implementações realizadas neste trabalho, seja possível recuperar o conhecimento necessário ao apoio a diagnóstico realizado pelos profissionais do Centro de Informações Toxicológicas, neste trabalho, representados pelo Núcleo de Santa Catarina (CIT/SC).

Na seção posterior serão feitos experimentos de modo a avaliar se a estrutura proposta possui viabilidade de utilização e também identificar qual a melhor estrutura modular se aplica ao contexto.

5. RESULTADOS

5.1. AVALIAÇÃO

A avaliação de desempenho do mecanismo desenvolvido com o presente estudo foi feita através do paradigma GQM (*Goal, Query, Metrics*) (Basili, 1988). Seguindo esta prática, foram determinados os parâmetros a seguir:

5.1.1. Definição

- **Objetivos**

Os objetivos desta avaliação são verificar a eficácia/eficiência do motor de busca criado juntamente com a ontologia implementada neste estudo, bem como definir qual a melhor abordagem tem a melhor aplicabilidade para recuperar conhecimento relacionado ao contexto de toxicologia clínica.

- **Questões**

Para medir a eficácia/eficiência do motor de busca, foram formuladas duas questões sobre sua performance:

- Qual a proporção de resultados relevantes em relação aos resultados obtidos em cada consulta realizada no experimento?
- Qual a proporção de resultados relevantes obtidos em relação ao total de resultados relevantes em cada experimento?

Concomitantemente, com o intuito de comparar e eleger a melhor estrutura de recuperação de conhecimento relacionada ao contexto de toxicologia clínica utilizando as abordagens construídas durante este estudo, a questão a ser respondida é a seguinte:

- Qual o melhor método e estrutura de conhecimento a ser utilizada em Sistemas de Recuperação de Conhecimento para Centros de Informações Toxicológicas?

Finalmente, para possibilitar a percepção sobre o aumento da produtividade com a utilização da ferramenta, devemos responder a questão:

- Houve melhora na performance com a utilização do mecanismo de busca proposto?

- **Métricas**

As métricas utilizadas neste estudo a fim de atender os requisitos da GQM são Precision e Recall (Salton e Mcgill, 1983), *Average Precision* – que é definida por Buckley e Voorhees (2000) como sendo "*medida de avaliação mais apropriada para mecanismos de recuperação de uso geral*" – além da métrica *Precision at Ten* (P@10) (Van Rijsbergen, 1974) para avaliar a precisão considerando os dez primeiros registros. Para a definição de performance geral com relação a questão levantada sobre o melhoramento na performance das consultas em relação aos métodos previamente utilizados no contexto dos CIT's, a métrica utilizada foi a de tempo em segundos.

As fórmulas destas métricas são definidas a seguir (Olson e Delen, 2008):

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{AveP} = \frac{\sum_{r=1}^{N} (P(r) \times \text{rel}(r))}{\text{number of relevant documents}}$$

Figura 41 - Métricas para avaliação de sistemas de Recuperação de Informação

5.1.2. Estudo de Caso

Neste estudo de caso foram realizados 50 experimentos trazendo como domínio as classes de substâncias constantes na ontologia em Toxicologia Clínica previamente ilustrada. Os tópicos foram selecionados por profissionais na área Toxicológica a fim de garantir a imparcialidade na avaliação e escolha dos termos para os testes. Foram utilizadas as principais classes de toxicologia/farmacologia definidas em (Dias *et al.*, 2001; Lüllmann e Langeloh, 2008) e inclusas duas substâncias para cada classe no set de testes.

O processo de seleção dos termos relevantes foi desenvolvido utilizando o método "*TREC 2002 Question Answering*" (Ellen, 2002) como parâmetro. Nos eventos TREC (Text REtrieval Conference) este método é utilizado baseado em um conjunto de questões previamente definido, para que cada tópico relacionado na marcação da base de

dados seja avaliado com base nestas questões. Neste estudo o método foi adaptado, devido ao fato de que o domínio do teste é uniforme, ou seja, diz respeito apenas a toxicologia clínica. Desta forma, a questão elaborada para avaliação é "a substância X é um valor relevante quando efetuada uma pesquisa pelo tópico Y?". Esta avaliação foi feita por um profissional da área toxicológica convidado para a pesquisa em cada um dos 50 tópicos selecionados por ele.

Após a base de conhecimento ter sido anotada, foram efetuados experimentos com os tópicos relacionados no motor de busca, com o intuito de fazer os cálculos de precisão e revocação do sistema. Para cada tópico foram computados os resultados em *Precision*, *Recall*, *Average Precision* e P@10. A análise e interpretação dos resultados são mostradas na seção seguinte.

5.2. ANÁLISE E INTERPRETAÇÃO

Nesta seção, serão demonstrados e analisados os resultados baseados nos critérios previamente apresentados, respondendo as questões baseadas nas médias dos resultados obtidos através dos experimentos.

Iniciando pela elucidação das questões 1 e 2, foram feitos experimentos usando as métricas básicas de *precision* e *recall* aplicadas a cada tópico. Os experimentos foram repetidos para cada método construído durante o estudo e foram comparados. Recapitulando, os métodos são:

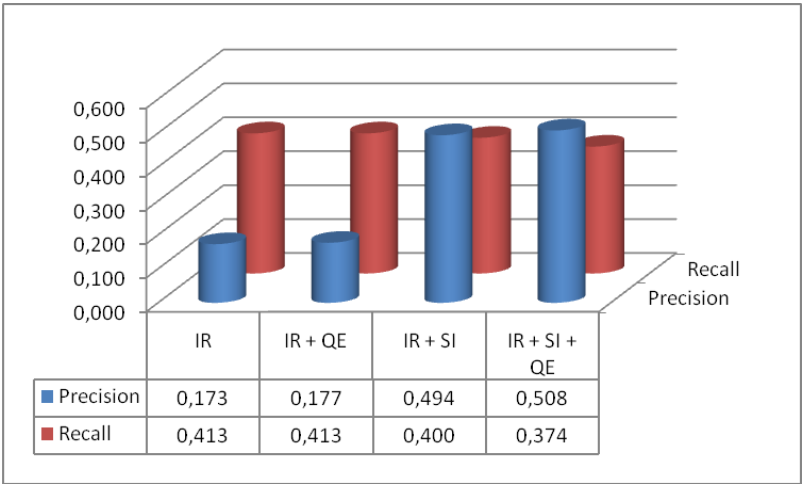
- Método tradicional de Recuperação de Informação (Information Retrieval - IR);
- IR com Expansão de Consulta;
- IR com Aprimoramento Semântico;
- IR + Expansão de Consulta + Aprimoramento Semântico.

Todos os experimentos foram efetuados utilizando uma única base de Conhecimento construída/descrita previamente. Na primeira fase do experimento, os testes foram efetuados para cada um dos tópicos selecionados pelo profissional toxicologista. As especificações e resultados produzidos nesta fase estão discriminados por método e descritos a seguir:

- IR tradicional: neste experimento, foram efetuados testes com os tópicos selecionados utilizando técnicas básicas

de Recuperação do Conhecimento, intrínsecas no Apache Solr, e retornaram como média de *Precision* (precisão) 0,173 e apresentaram média de *Recall* (revocação) de 0,413;

- IR + Expansão de Consulta (QE): nesta fase de testes, o mecanismo tradicional recebeu o auxílio do mecanismo de expansão de consultas. A média registrada para o *precision* apontou um pequeno avanço, resultando em 0,177. Para este caso, o recall permaneceu em 0,413;
- IR + Aprimoramento Semântico (Semantic Improvement – SI): com o aprimoramento semântico, os testes mostraram uma média de precision de 0,494 e média de recall de 0,400. Houve um avanço considerável neste método, se comparado com as técnicas tradicionais de recuperação de informação. Faz-se necessário registrar uma redução no *Precision* em 0,013;
- IR + SI + QE: neste experimento, foram adicionadas todas as técnicas desenvolvidas neste estudo na arquitetura do motor de busca. A média de *Precision* foi de 0,508, enquanto que a média de recall foi de 0,374.



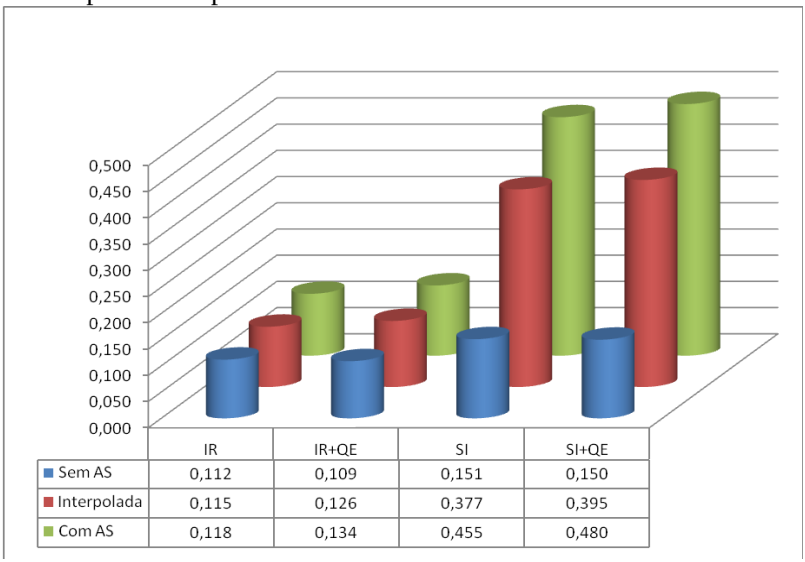
	IR	IR + QE	IR + SI	IR + SI + QE
Precision	0,173	0,177	0,494	0,508
Recall	0,413	0,413	0,400	0,374

Figura 42- Gráfico de representação dos resultados obtidos pelos diferentes métodos utilizando a as métricas Precision e Recall

Como pode ser visto, a estrutura sem Aprimoramento Semântico proveram resultados não satisfatórios, visto que obtiveram

como resultado poucos tópicos relevantes em proporção. Nos primeiros testes com o Aprimoramento Semântico, houve um avanço significativo na performance relacionado à precisão, apresentando uma leve queda na revocação.

Para responder a terceira questão proposta anteriormente, foram utilizadas as métricas *Average Precision* e P@10 para cada método implementado, conforme previamente citado. Após feitos os primeiros experimentos com todos os métodos de Recuperação, foram obtidos os resultados que podem ser visualizados nas Figura 42 e Figura 43, que serão expostos e explanados mais adiante.



	IR	IR+QE	SI	SI+QE
Sem AS	0,112	0,109	0,151	0,150
Interpolada	0,115	0,126	0,377	0,395
Com AS	0,118	0,134	0,455	0,480

Figura 43 - Gráfico de representação dos resultados obtidos pelos diferentes métodos e fragmentos da base de conhecimento utilizando a métrica Average Precision

Estes experimentos foram feitos utilizando a mesma base de conhecimento que os testes anteriores. Assim como nos testes com *Precision* e *Recall*, foram utilizados os mesmos tópicos para cada método. As especificações e resultados dos testes podem ser vistos a seguir:

- IR tradicional: foram efetuados os testes nos 50 tópicos utilizando somente técnicas básicas de Recuperação de Conhecimento, que são nativos do Apache Solr. A média de *Average Precision* retornada foi 0,115 com um desvio padrão registrado de 0,088, enquanto que o P@10 foi de 0,28;

- IR + QE: os testes foram efetuados com auxílio do mecanismo de Expansão de Consulta. A média registrada de *Average Precision* foi de 0,126 com um desvio padrão de 0,113, enquanto o P@10 foi de 0,27;
- IR + SI: com o Aperfeiçoamento Semântico, o *Average Precision* confirmou o avanço de performance já registrado pelo *Precision* e *Recall*, resultando em AP de 0,377 e P@10 de 0,55;
- IR + SI + QE: como nos experimentos com as métricas de *Precision* e *Recall*, neste caso também foram concatenados todos os métodos de aperfeiçoamento no motor de busca para que fossem feitos os testes. As métricas AP e P@10 apontaram que houve progresso em relação ao método anterior utilizado (IR + SI), resultando em 0,395 e 0,57 respectivamente, com desvio padrão de 0,35. Pode-se observar que a métrica P@10 aponta que o primeiro método tem uma média melhor com a base sem anotações semânticas. Isto ocorre porque, na base externa de aprimoramento semântico, existem resultados coincidentes com os esperados na pesquisa. Também existe o fator a ser considerado que, a busca sintática feita nos campos pré-definidos apresenta mais exatidão, quando efetuada por palavras-chave.

Após todos os experimentos com as métricas AP e P@10 tendo sido efetuados, foi possível observar que a média de desvio padrão obtida nos testes foi alta. Foi descoberto que o motivo desta discrepância se dava pela existência de duas populações distintas de tópicos na base de conhecimento: tópicos que continham anotações semânticas (*semantic annotations* – SA) e tópicos que não possuíam estas anotações.

No intuito de resolver esta questão, as populações foram divididas em duas sub-populações distintas a partir desta característica (com e sem anotações semânticas). Ao final desta divisão, foram obtidos 36 registros com SA e 14 registros sem SA, e então foram novamente calculados os desvios padrões separadamente.

Com a segunda parcela de registros (base sem anotações), o desvio padrão foi grandemente reduzido, devido a homogeneização da população. Com a segunda parte dos registros, o desvio ainda continuou

alto. A causa deste acontecimento será discutido na próxima seção, bem como serão discutidos os resultados apresentados.

Para responder a quarta questão proposta nesta validação, nós entrevistamos um profissional em toxicologia clínica habituado com a prática atual exercida no CIT de Santa Catarina, e lhe propusemos uma simulação de uso da ferramenta em comparação com os métodos que seriam utilizados por ele em uma situação de emergência. A comparação dos resultados foram computadas e comparadas em medida de tempo, ou seja, tempo gasto desde o momento em que se inicia uma consulta até o momento em que se chega a uma resposta válida.

Os testes foram realizados em 10 dos 50 registros previamente estudados nos experimentos anteriores, levando em consideração as respostas válidas de acordo com os itens relevantes previamente selecionados.

Optou-se pela utilização de recursos disponíveis na web como comparação com o mecanismo de busca atual, ou seja, foram utilizadas bases de dados sobre toxicologia clínica em mecanismos de buscas já existentes, dispostos em um padrão semelhante ao do mecanismo proposto neste estudo.

Observou-se que, em todos os casos, o retorno medido em tempo foi mais rápido no mecanismo de busca proposto, o que se explica pelos seguintes fatores:

1. A base de conhecimento contém o conhecimento necessário as buscas persistido em uma base local, enquanto que os mecanismos encontrados na web fazem uma varredura em toda a web, a fim de encontrar o que se procura;

2. O contexto de procura delimitado pelo motor de busca proposto é previamente definido, limitando-se apenas ao domínio de toxicologia clínica.

Estas observações nos remetem a uma ressalva a ser feita com relação à utilização do mecanismo de busca: este mecanismo tem bom desempenho nas consultas cujo conhecimento necessário a um atendimento já esteja contido na base de conhecimento em questão. Com relação ao comparativo de desempenho em medida de tempo, a tabela com os dados dos testes computados está disponível no anexo 4 e a ilustração na Figura 44.

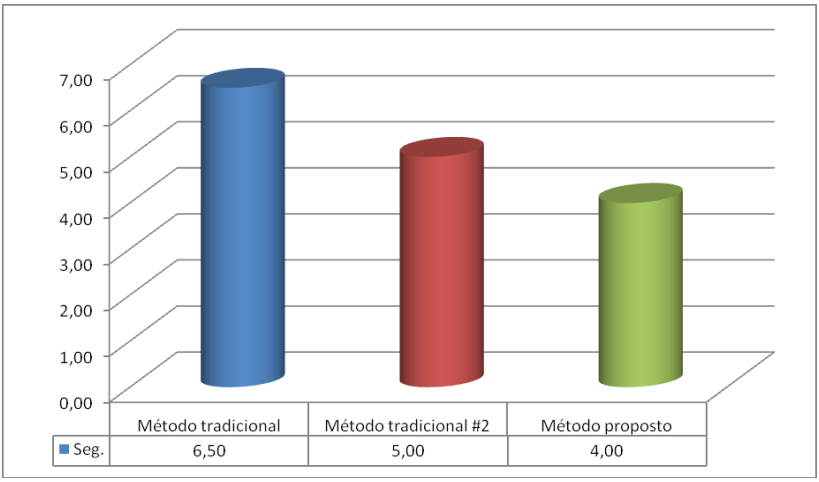


Figura 44 - Gráfico comparativo entre os mecanismos de busca

5.3. ANÁLISE E DISCUSSÃO DOS RESULTADOS

5.3.1. Ambiente experimental

Os experimentos realizados para a avaliação foram efetuados utilizando dois equipamentos de *hardware* distintos. Para a instalação da estrutura de *software* desenvolvida para a proposta, foi utilizada um computador portátil (*notebook*) com a seguinte configuração:

- Processador: AMD Turion X2 Dual Core Mobile RM-70 – 2.0 GHz;
- Memória: 4.0 GB;
- Disco Rígido: 120 GB e
- Sistema operacional: Linux Ubuntu 9.04 – 64 bits.

Para a manipulação da interface desenvolvida para este sistema, foi utilizado um equipamento com monitor de sensibilidade ao toque (*touch screen*), com a seguinte configuração:

- Processador: Intel® Core™ i3 330M 2,13 GHz, DMI 2,5 GT/s
- Memória: DDR3 de 4 GB
- Disco rígido 3G SATA 1 TB (5400 rpm)
- Sistema Operacional: Windows® 7 Home Premium 64 bit

5.3.2. Recursos humanos envolvidos e avaliação de utilização

Para a avaliação dos resultados, esta foi dividida em duas etapas: usabilidade e performance. Foram convidados à participação e colaboração nos testes dois profissionais envolvidos com o Centro de Informações Toxicológicas de Santa Catarina – CIT/SC, visto que o primeiro deles já exercia suas atividades externamente à instituição, enquanto que o colaborador com a análise de interface estava diretamente alocado no CIT.

O primeiro profissional é farmacêutico formado pela Universidade Federal de Santa Catarina, tendo atuado no CIT e em virtude disso, sendo conhecedor da metodologia de atendimento implementada no Centro, além de possuir extenso conhecimento sobre agentes tóxicos, o que possibilitou aferir os resultados obtidos com melhor profundidade de análise.

Embora o foco principal da avaliação dos resultados seja predominantemente voltado à performance, a análise de utilização realizada pelo colaborador atuante no CIT - além de depoimento de outros profissionais que observaram a utilização da proposta – mostrou-se bastante positivo com relação à usabilidade do sistema, observando pontos de melhoramento considerável na atividade de pesquisa e vinculação dos agentes tóxicos à ficha informatizada desenvolvida especialmente para este fim.

5.3.3. Avaliação dos Resultados e Implicações

Os resultados dos experimentos realizados mostraram basicamente que houve um incremento na performance do mecanismo de busca a cada novo método aplicado. Embora o aperfeiçoamento considerado na inclusão do segundo método (a inclusão da expansão da consulta) não tenha sido plenamente satisfatório, deve ser levado em conta o fato de que a base de conhecimento utilizada nos experimentos foi um recorte da base plena construída neste estudo, e desta forma, em testes com a base completa, há a possibilidade da variação dos resultados e o aperfeiçoamento com a expansão da consulta pode ser mais aparente.

Correlacionando o primeiro método com o terceiro método de recuperação (com o aperfeiçoamento semântico), foi possível observar um aumento considerável na performance do motor de busca. Isso ocorre devido ao fato de que o motor de busca faz o relacionamento entre os conceitos através da ação da camada de Aperfeiçoamento

Semântico juntamente com as anotações semânticas contidas na base de conhecimento.

E no quarto método este avanço foi ainda mais visível, mostrando um grande aumento registrado pela métrica *precision* e uma pequena queda registrada pela métrica *recall*, o que pode ser considerado normal em se tratando de sistemas de Recuperação. Com o aperfeiçoamento semântico, o mecanismo de expansão de consulta teve um melhor desempenho do que nos experimentos anteriores com este método. A explicação para este fenômeno se dá pela maior cobertura relacional entre a sinonímia proporcionada pela expansão de consulta em conjunto com a gama de itens relacionados mapeados pelas anotações semânticas.

A avaliação dos experimentos através da métrica *Average Precision* também mostra que, em geral, o quarto método (a concatenação de todos os métodos implementados neste estudo) produz melhores resultados, mas existem fatores a serem considerados, que concretizam a afirmação: "bases de conhecimento sem anotações semânticas podem produzir resultados insatisfatórios quando utilizadas em conjunto com camadas de Aperfeiçoamento semântico". Embora esta afirmação possa parecer óbvia, esperava-se que os resultados sem o ajuste semântico fossem no mínimo similares aos resultados produzidos pelos métodos sem aperfeiçoamento semântico, porém, em alguns casos a camada SI, quando utilizada em tópicos sem anotações semânticas, ocultou respostas válidas, como pode ser visto nos tópicos de números 10, 18 e 30 da tabela disponibilizada no apêndice. A camada semântica mostrou-se incompatível com o mecanismo de recuperação do Apache Solr quando os tópicos relacionados não continham anotações semânticas.

Assim, foi comprovado que para se utilizar com bom grau de confiança uma camada de aperfeiçoamento semântico em uma aplicação com o tipo de estrutura abordada neste estudo, é necessário se ter uma base de conhecimento compatível (devidamente anotada semanticamente), a fim de evitar problemas como os que foram demonstrados nos tópicos supracitados.

5.3.4. Ameaças à Validação

Existem alguns fatores que podem representar ameaças a validação neste estudo. Um destes fatores diz respeito ao recorte da base de conhecimento feito para possibilitar a realização dos experimentos, a

fim de permitir a demarcação dos resultados relevantes e irrelevantes para os tópicos em toda a base de conhecimento, recorte feito devido ao tempo e recursos disponíveis para tal.

A base do DeCS foi utilizada pelo fato de que possui em língua portuguesa: conceitos, descritores, sinônimos, além de possuir também sinônimos na língua inglesa para possibilitar as conexões com os conceitos da base de conhecimento. Assim sendo, os resultados podem ter alterações no momento em que for utilizada a base de conhecimento na íntegra, pois podem haver mais conexões entre sinônimos/descritores do DeCS e ocorrências na base de conhecimento.

Outra ameaça a ser discutida é a subjetividade da avaliação no momento da seleção dos resultados relevantes para os tópicos. Para esta avaliação utilizando a questão previamente definida no GQM, foi convidado um profissional experiente na área de toxicologia clínica e apto ao desenvolvimento deste trabalho, porém, esta avaliação é baseada em literatura e conhecimento prévio do avaliador, o que pode acarretar em pequenos desacordos entre a literatura utilizada pelo avaliador e o conteúdo da base de conhecimento, o que pode vir a alterar os resultados se examinados por outro avaliador.

5.3.5. Inferências

Após o trabalho completado, é possível inferir que os resultados serão melhores quando o mecanismo for usado com a base de conhecimento na íntegra. Isto porque, vai haver maior conteúdo que seja coincidente com o pesquisado quando utilizada a camada de aperfeiçoamento semântico e a expansão de consulta, mas como visto anteriormente, isto somente vai acontecer se esta base de conhecimento contiver as anotações semânticas.

Possivelmente haverá um aumento do *recall* devido ao incremento da performance entre os relacionamentos entre itens relevantes, que vão aparecer nos primeiros resultados apresentados pelo mecanismo.

Com a integração das bases de conhecimento adicionais a base estrutural previamente construída neste estudo, será possível retornar mais registros coincidentes na língua portuguesa.

Outro fator a ser levado em conta é que, usando sentenças mais completas (não só palavras-chave, como foram efetuados os testes), a possibilidade de incremento na acurácia dos resultados é considerável, devido ao grande número de parâmetros a serem utilizados na pesquisa.

5.3.6. Comparativo: Trabalhos relacionados VS. Proposta

Nesta seção será reproduzida a tabulação dos trabalhos relacionados abordados previamente, ilustrando juntamente as funcionalidades cobertas pela proposta apresentada neste trabalho.

Autor	IDX	SM	QE	IBDH
(Berners-Lee *et al.*, 2001)		x		x
(Rocha *et al.*, 2004)		x		
(Wang e Jhuo, 2009)		x		
(Ilyas *et al.*)		x		x
(Bernstein *et al.*, 2006)				
(Tang *et al.*)	x			
(Minack *et al.*, 2008)	x			
(Heese *et al.*, 2007)	x			
(Díaz-Galiano *et al.*, 2009)	x		x	
(2007)			x	
(Yoo e Choi, 2010)			x	
PROPOSTA	X	X	X	X

A tabela demonstra que a abordagem da proposta deste trabalho visa englobar um conjunto de técnicas para contemplar o maior número de possibilidades de colaboração a fim de melhorar a performance da busca.

6. CONCLUSÃO

Neste trabalho foram realizados experimentos com aplicações para auxiliar os profissionais dos Centros de Informações Toxicológicas na pesquisa sobre agentes tóxicos no intuito de facilitar este processo em atendimentos de urgência. O problema o qual este estudo propôs solução diz respeito basicamente à homogeneidade e dispersão do conhecimento necessário para o referido atendimento de urgência.

Os objetivos específicos estabelecidos a partir do problema foram atingidos através da construção de uma arquitetura modular composta por artefatos e técnicas adaptadas para este contexto.

O primeiro objetivo específico diz respeito à criação de um artefato que contivesse o conhecimento necessário à demanda de consultas no processo de atendimento emergencial. Objetivo este atingido através da modelagem de uma ontologia que regeria estruturalmente a posterior construção da base de conhecimento, base esta que contém as instâncias relacionadas aos itens de agentes tóxicos importados de fontes previamente estudadas por profissionais e descritas no processo como sendo fontes de conhecimento relevantes. Esta construção da base de conhecimento foi auxiliada pelo desenvolvimento de um mecanismo capaz de gerenciar os diferentes formatos de persistência do conhecimento e posterior armazenamento na nova disposição homogênea resultante.

Para fazer o acesso ao conhecimento constante no novo artefato gerado, fazia-se necessária a estruturação de um mecanismo capaz de lidar com o novo conteúdo gerado. Para isso foram estudadas técnicas de ferramentas de indexação e recuperação, expansão de consulta, técnicas de busca semântica, fonética, entre outras. Estas técnicas e mecanismos foram integrados em um único motor de busca implementado para a manipulação da base de conhecimento. Devido ao grande número de pesquisas efetuadas nesta área, haviam várias combinações possíveis para utilização de diferentes modelos estruturais para o mecanismo. Desta forma, foram avaliadas separadamente cada uma das combinações apresentadas durante o estudo através de métricas específicas usuais a este tipo de trabalho. Através desta avaliação, o objetivo específico relacionado a este item foi atendido com a utilização da estrutura melhor avaliada através das métricas.

Foi possível perceber através da análise comparativa entre os métodos previamente utilizados (feita com a ajuda de um profissional em toxicologia clínica) que o desempenho geral obtido através da utilização da ferramenta desenvolvida neste estudo apresentou

resultados melhores do que as anteriores em decorrência das especificidades já discutidas na seção anterior.

Também foi implementada uma interface adaptada para utilização com equipamentos com tecnologia de sensibilidade a toque (*touch screen*). A avaliação de ergonomia disponibilizada através da utilização desta interface foi sugerida como trabalhos futuros, como poderá ser visto na seqüência.

Observou-se ao final das avaliações realizadas que este mecanismo pode ser utilizado para fins genéricos com adaptações de interface e bases de conhecimento principal e auxiliares. No aspecto da utilização de aperfeiçoamento semântico ficou claro que, para utilização da estrutura do motor de busca na íntegra, a base de conhecimento deve possuir as anotações semânticas para não haver problemas na recuperação.

6.1. SUGESTÕES PARA TRABALHOS FUTUROS

Para a possibilitar a completude e o melhor desempenho do protótipo implementado neste estudo, transformando este em um elemento de utilização plena para os fins para que foi modelado, sugere-se como trabalhos futuros os itens que seguem:

- Aperfeiçoamento do módulo de Aprimoramento Semântico para suportar tecnologias de busca semântica com sensibilidade a contexto;
- Avaliação de novas possibilidades de percursos de navegação na árvore do DeCS no módulo de expansão de consulta;
- Trabalhos de avaliação sobre a usabilidade da interface *touch screen* através de métodos heurísticos apropriados, sugerindo-se a utilização de técnicas similares às de análise de Nielsen (Nielsen, 1994);
- Criação de um sistema para automatização da atualização da base de conhecimento, que manipule os *crawlers* já desenvolvidos guiado por uma ontologia específica, definindo destinos de busca por contexto.

REFERÊNCIAS

ABEL, M. Sistemas de conhecimento. *Notas de Aula. Porto Alegre, Instituto de Informática da UFRGS* [S.I.], 2002.

AGOSTI, M.; BONFIGLIO-DOSIO, G.; FERRO, N. A historical and contemporary study on annotations to derive key features for systems design. *International Journal on Digital Libraries* [S.I.], v. 8, n. 1, p. 1-19, 2007.

AGROFIT. Sistema de Agrotóxicos Fitossanitários. In: HTTP://WWW.APACHE.ORG, D. E. (Ed.)2010.

AIJUAN, D.; HONGLIN, L. Ontology-based information integration in virtual learning environment. In: Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on. 2005. p.762-765. Disponível em:<10.1109/WI.2005.108>. Acesso em.

ALAKO, B. T.; VELDHOVEN, A.; VAN BAAL, S. *et al.* CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics* [S.I.], v. 6, p. 51, 2005.

ALAVI, M.; LEIDNER, D. Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS quarterly* [S.I.], v. 25, n. 1, p. 107-136, 2001.

ALMEIDA, M.; BAX, M. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. *Ci. Inf* [S.I.], v. 32, n. 3, p. 7-20, 2003.

ALONSO-CALVO, R.; MAOJO, V.; BILLHARDT, H. *et al.* An agent- and ontology-based system for integrating public gene, protein, and disease databases. *Journal of Biomedical Informatics* [S.I.], v. 40, n. 1, p. 17-29, 2007.

ALPI, K. Expert searching in public health. *Journal of the Medical Library Association* [S.I.], v. 93, n. 1, p. 97, 2005.

ANDERSON, C. A.; COPESTAKE, P. T.; ROBINSON, L. A specialist toxicity database (TRACE) is more effective than its larger, commercially available counterparts. *Toxicology* [S.I.], v. 151, n. 1-3, p. 37-43, 2000.

ANGELE, J.; FENSEL, D.; LANDES, D. *et al.* Developing Knowledge-Based Systems with MIKE. *Automated Software Engineering* [S.I.], v. 5, n. 4, p. 389-418, 1998.

ANVISA. Agência Nacional de Vigilância Sanitária. v. 2010, n. 21/06/2010, 2009. Disponível em: http://portal.anvisa.gov.br/.

APACHE. Apache http server project. *Disponível em: http://www.apache.org*. v. 2009. n. 01/09/2009.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern information retrieval*. Addison-Wesley Reading, MA, 1999.

BASILI, V. R. The TAME project: towards improvement-oriented softwareenvironments. *IEEE Transactions on Software Engineering* [S.I.], v. 14, p. 758-773, 1988.

BATEMAN, D.; GOOD, A.; KELLY, C. *et al.* Web based information on clinical toxicology for the United Kingdom: uptake and utilization of TOXBASE in 2000. *British journal of clinical pharmacology* [S.I.], v. 54, n. 1, p. 3-9, 2002.

BELLATRECHE, L.; DUNG, N. X.; PIERRA, G. *et al.* Contribution of ontology-based data modeling to automatic integration of electronic catalogues within engineering databases. *Computers in Industry* [S.I.], v. 57, n. 8-9, p. 711-724, 2006.

BERMAN, L.; CULLEN, M.; MILLER, P. Automated integration of external databases: a knowledge-based approach to enhancing rule-based expert systems. American Medical Informatics Association, 1992. p.227.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific american* [S.I.], v. 284, n. 5, p. 28-37, 2001.

BERNSTEIN, A.; KAUFMANN, E.; KAISER, C. *et al.* Ginseng: A guided input natural language search engine for querying ontologies. Citeseer, 2006.

BIFFL, S.; SUNINDYO, W. D.; MOSER, T. Semantic Integration of Heterogeneous Data Sources for Monitoring Frequent-Release Software Projects. In: Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International Conference on. 2010. p.360-367. Disponível em:<10.1109/CISIS.2010.58>. Acesso em.

BIREME. DeCS/VMX. n. 15 fev 2009, 2010. Disponível em:<http://decs.bvs.br/vmx.htm>. Acesso em: 10/10/2009.

BODENREIDER, O.; BURGUN, A. Aligning knowledge sources in the UMLS: methods, quantitative results, and applications. Ios Pr Inc, 2004. p.327.

BOFF, L. Processo Cognitivo de Trabalho de Conhecimento: Um Estudo Exploratório sobre O Uso da Informação no Ambiente de Análise de Investimentos. *Universidade Federal do Rio Grande do Sul* [S.I.], 2000.

BOGDANOWICZ, M.; BAILEY, E. The value of knowledge and the values of the new knowledge worker: generation X in the new economy. *Training* [S.I.], v. 4, n. 2002, p. 125-129, 2003.

BONIFATI, A.; CHANG, E.; HO, T. *et al.* Schema mapping and query translation in heterogeneous P2P XML databases. *The VLDB Journal* [S.I.], v. 19, n. 2, p. 231-256, 2010.

BORST, W. Construction of engineering ontologies. *University of Tweenty. Enschede, NL-Centre for Telematica and Information Technology* [S.I.], 1997.

BREUKER, J.; VAN DE VELDE, W. *CommonKADS library for expertise modelling: reusable problem solving components*. IOS press, 1994.

BROEKSTRA, J.; KAMPMAN, A.; VAN HARMELEN, F. Sesame: A generic architecture for storing and querying rdf and rdf schema. *The Semantic Web—ISWC 2002* [S.I.], p. 54-68, 2002.

BUCKLEY, C.; VOORHEES, E. Evaluating evaluation measure stability. ACM, 2000. p.40.

CABRAL, R.; ANDRADE, R.; SAVARIS, A. *et al.* Plataforma de Gerência do Conhecimento Aplicada em um Ambiente de Toxicologia Clínica e Toxicovigilância. *In: Congresso Brasileiro de Informática na Saúde, 2008, Campos do Jordão - SP. Congresso Brasileiro de Informática na Saúde. São Paulo*: SBIS, 2008.

CABRAL, R. B.; ANDRADE, R.; JUNIOR, C. L. B. *et al.* Semantic Information Indexing and Retrieval on Patient Medical Data. *8th International Information and Telecommunication Technologies Symposium*. v. 8. Florianópolis: Fundação Barddal de Educação e Cultura, 2009. p. 171-174.

CAI, D.; RIJSBERGEN, C. J. V.; JOSE, J. M. Automatic query expansion based on divergence. *Proceedings of the tenth international conference on Information and knowledge management*. Atlanta, Georgia, USA: ACM, 2001. p. 419-426.

CARDOSO, O. Recuperação de informação. *Semana de Ciência da Computação. UFLA, Lavras: MG* [S.I.], 2000.

CARROLL, J.; DICKINSON, I.; DOLLIN, C. *et al.* Jena: implementing the semantic web recommendations. ACM New York, NY, USA, 2004. p.74-83.

CHANDRASEKARAN, B.; JOSEPHSON, J.; BENJAMINS, V. What are ontologies, and why do we need them? *IEEE Intelligent systems* [S.I.], v. 14, n. 1, p. 20-26, 1999.

CHANG, W.; SHEIKHOLESLAMI, G.; WANG, J. *et al.* Data resource selection in distributed visual information systems. *Knowledge and Data Engineering, IEEE Transactions on* [S.I.], v. 10, n. 6, p. 926-946, 1998.

CHANG, Y.; CHEN, S.; LIAU, C. A New Query Expansion Method for Document Retrieval Based on the Inference of Fuzzy Rules. *Journal-Chinese Institute of Engineers* [S.I.], v. 30, n. 3, p. 511, 2007.

CHU, W. W.; CÁRDENAS, A. F.; TAIRA, R. K. KMeD: A knowledge-based multimedia medical distributed database system. *Information Systems* [S.I.], v. 20, n. 2, p. 75-96, 1995.

CHUNG, C.-W. DATAPLEX: an access to heterogeneous distributed databases. *Commun. ACM* [S.I.], v. 33, n. 1, p. 70-80, 1990.

CIMINO, J.; LI, J.; GRAHAM, M. *et al.* Use of online resources while using a clinical information system. American Medical Informatics Association, 2003.

CLANCEY, W. J. The epistemology of a rule-based expert system --a framework for explanation. *Artificial Intelligence* [S.I.], v. 20, n. 3, p. 215-251, 1983.

CLANCEY, W. J. Heuristic classification. *Artificial Intelligence* [S.I.], v. 27, n. 3, p. 289-350, 1985.

CLANCEY, W. J. The Knowledge Level Reinterpreted: Modeling How Systems Interact. *Machine Learning* [S.I.], v. 4, n. 3, p. 285-291, 1989.

DARMONI, S. J.; NEVEOL, A.; RENARD, J. M. *et al.* A MEDLINE categorization algorithm. *BMC Med Inform Decis Mak* [S.I.], v. 6, p. 7, 2006.

DAVID, J.; KRIVINE, J.; SIMMONS, R. *Second generation expert systems*. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 1993.

DAVIS, R.; BUCHANAN, B.; SHORTLIFFE, E. Production rules as a representation for a knowledge-based consultation program. *Artificial Intelligence* [S.I.], v. 8, n. 1, p. 15-45, 1977.

DIAS, A.; WELFER, D.; D'ORNELLAS, M. JENA: UMA FERRAMENTA PARA DESENVOLVER COMUNIDADES VIRTUAIS DE PESQUISA CIENTÍFICA. *Revista do CCEI* [S.I.], v. 8, p. 14, 2004.

DIAS, M.; CAMPOLINA, D.; GUERRA, S. *et al.* Toxicologia na Prática Clínica. *ANDRADE FILHO, A. Toxicologia na Prática Clínica. 1ed. Belo Horizonte: Folium* [S.I.], p. 155-165, 2001.

DÍAZ-GALIANO, M.; MARTÍN-VALDIVIA, M.; UREÑA-LÓPEZ, L. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in Biology and Medicine* [S.I.], v. 39, n. 4, p. 396-403, 2009.

DING, Y.; EMBLEY, D.; LIDDLE, S. Automatic creation and simplified querying of semantic web content: An approach based on information-extraction ontologies. *The Semantic Web–ASWC 2006* [S.I.], p. 400-414, 2006.

DOERR, M.; IORIZZO, D. The dream of a global knowledge network\—A new approach. *J. Comput. Cult. Herit.* [S.I.], v. 1, n. 1, p. 1-23, 2008.

DUARTE, E. Classificação facetada: um olhar sobre a construção de estruturas semânticas Faceted classification: a look at the construction of semantic structures p. 46-58. *Revista Digital de Biblioteconomia e Ciencia da Informação* [S.I.], v. 7, n. 2, 2010.

ELLEN, M. Overview of the TREC 2002 Question Answering Track. 2002.

EUZENAT, J.; ISAAC, A.; MEILICKE, C. *et al.* Ontology Matching. 2007. Disponível em:<http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/3540496114>.

EYAL, A.; MILO, T. Integrating and customizing heterogeneous e-commerce applications. *The VLDB Journal* [S.I.], v. 10, n. 1, p. 16-38, 2001.

FARIAS, R.; DE MATTOS, M.; WALESKA, P. *et al.* Ontologia para a Gestão do Conhecimento em Saúde por meio da Metodologia Methontology.

FENSEL, D.; VAN HARMELEN, F. A comparison of languages which operationalize and formalize KADS models of expertise. *The Knowledge Engineering Review* [S.I.], v. 9, n. 02, p. 105-146, 1994.

FERNANDEZ, M.; GOMEZ-PEREZ, A.; JURISTO, N. Methontology: From ontological art towards ontological engineering. 1997. p.33–40.

FONGER, G.; STROUP, D.; THOMAS, P. *et al.* TOXNET: A computerized collection of toxicological and environmental health information. *Toxicology and Industrial Health* [S.I.], v. 16, n. 1, p. 4, 2000.

GEHANNO, J. F.; PARIS, C.; THIRION, B. *et al.* Assessment of bibliographic databases performance in information retrieval for occupational and environmental toxicology. *Occup Environ Med* [S.I.], v. 55, n. 8, p. 562-6, Aug 1998.

GENNARI, J.; MUSEN, M.; FERGERSON, R. *et al.* The evolution of Protégé: an environment for knowledge-based systems development. *International Journal of Human-Computer Studies* [S.I.], v. 58, n. 1, p. 89-123, 2003.

GONG, P.; FENG, D.; LIM, Y. S. An Intelligent Middleware for Dynamic Integration of Heterogeneous Health Care Applications. In: Multimedia Modelling Conference, 2005. MMM 2005. Proceedings of the 11th International. 2005. p.198-205.

GOSPODNETIC, O.; HATCHER, E. *Lucene in Action (In Action series)*. Manning Publications, 2004.

GREENGRASS, E. Information retrieval: A survey. *preservation*, v. 2, p.6, 2001. Disponível em:<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.1855>.

GRUBER, T. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies* [S.I.], v. 43, n. 5, p. 907-928, 1995.

GRUBER, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies*, v. 43, n. 5-6, p.907-928, 1995. Disponível em:<http://www.sciencedirect.com/science/article/B6WGR-45NJJDF-K/2/b47f5cb67315c76b60ac39f44e0a2cec>.

GRUNINGER, M.; FOX, M. Methodology for the Design and Evaluation of Ontologies. 1995.

GUARINO, N. Formal Ontology and Information Systems. In: Proceedings of FOIS'98, 6-8 de junho de 1988, Trento - Italia. Italia: IOS Press, 1998. p.3-15. Disponível em:<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.1776>. Acesso em: 01-fev-2010.

GUBANOV, M. N.; POPA, L.; HO, H. *et al.* IBM UFO repository: object-oriented data integration. *Proc. VLDB Endow.* [S.I.], v. 2, n. 2, p. 1598-1601, 2009.

GUIMARÃES, F. Utilização de ontologias no domínio B2C. *Mestrado em Informática, Pontifícia Universidade Católica do Rio de Janeiro* [S.I.], 2002.

GUPTA, A.; CONDIT, C.; QIAN, X. BioDB: An ontology-enhanced information system for heterogeneous biological information. *Data & Knowledge Engineering* [S.I.], v. In Press, Corrected Proof, 2010.

HA, J.; WEI, Y.; JIN, Y. Logistics Decision-making Support System Based on Ontology. In: Computational Intelligence and Design, 2008. ISCID '08. International Symposium on. 2008. p.309-312. Disponível em:<10.1109/ISCID.2008.128>. Acesso em.

HANSEN, M.; NOHRIA, N.; TIERNEY, T. What's your strategy for managing knowledge? *Knowledge management: critical perspectives on business and management* [S.I.], v. 77, n. 2, p. 322, 2005.

HART, P.; DUDA, R.; EINAUDI, M. PROSPECTOR—A computer-based consultation system for mineral exploration. *Mathematical Geology* [S.I.], v. 10, n. 5, p. 589-610, 1978.

HATCHER, E.; GOSPODNETIC, O. *Lucene in Action (In Action series).* Manning Publications, 2004. Disponível em:<http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/1932394281>. Acesso em: 12/10/2009.

HEESE, R.; LESER, U.; QUILITZ, B. *et al.* Index Support for SPARQL. *ESWC, Innsbruck, Austria* [S.I.], 2007.

HSDB, H. National Library of Medicine. *Toxicology Information Program, Washington, DC* [S.I.], 1991.

HUISMANS, J. International Register of Potentially Toxic Chemicals(IRPTC). *ECOTOXICOL. AND ENVIRON. SAFETY* [S.I.], v. 4, n. 4, p. 393-403, 1980.

ILYAS, Q.; KAI, Y.; TALIB, M. A Conceptual Architecture for Semantic Search Engine.

KENTON, C.; SCOTT, Y. B. MEDLINE searching and retrieval. *Med Inform (Lond)* [S.I.], v. 3, n. 3, p. 225-35, Sep 1978.

KIRYAKOV, A.; POPOV, B.; TERZIEV, I. *et al.* Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web* [S.I.], v. 2, n. 1, p. 49-79, 2004.

KITCHENHAM, B. Procedures for performing systematic reviews. *NICTA Technical Report 0400011T.1*, 2004. Disponível em:<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.122.3308&rep=rep1&type=pdf>. Acesso em: 10/10/10.

KNOX, K. A Researcher's Dilemma-Philosophical and Methodological Pluralism. *Electronic Journal of Business Research Methods* [S.I.], v. 2, n. 2, p. 119-128, 2004.

KNUBLAUCH, H.; FERGERSON, R.; NOY, N. *et al.* The Protégé OWL plugin: An open development environment for semantic web applications. *Lecture notes in computer science* [S.I.], p. 229-243, 2004.

KNUBLAUCH, H.; FERGERSON, R. W.; NOY, N. F. *et al.* The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications*The Semantic Web – ISWC 2004*, 2004. p. 229-243.

KOIVUNEN, M.; MILLER, E. W3C semantic web activity. *Semantic Web Kick-Off in Finland* [S.I.], p. 27–44, 2001.

KOLB, D. *Experiential learning: Experience as the source of learning and development*. Prentice Hall, 1984.

KONG, G.; XU, D.; YANG, J. Clinical decision support systems: a review on knowledge representation and inference under uncertainties. *International Journal of Computational Intelligence Systems* [S.I.], v. 1, n. 2, p. 159–167, 2008.

KOUFI, V.; MALAMATENIOU, F.; VASSILACOPOULOS, G. A system for the provision of medical diagnostic and treatment advice in home care environment. *Personal Ubiquitous Comput.* [S.I.], v. 14, n. 6, p. 551-561, 2010.

KURLYANDSKIY, B. A.; SIDOROV, K. K. History and current state of toxicology in Russia. *Toxicology* [S.I.], v. 190, n. 1-2, p. 55-62, 2003.

LANDRY, R.; AMARA, N.; PABLOS-MENDES, A. *et al.* The knowledge-value chain: a conceptual framework for knowledge translation in health. *Bulletin of the World Health Organization* [S.I.], v. 84, p. 597-602, 2006.

LASSILA, O.; SWICK, R. Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation 22 February 1999. *W3C-World Wide Web Consortium,[Online] http://www. w3. org/TR/REC-rdf-syntax* [S.I.].

LAWRENCE, D. W.; GUARD, A.; MEIER, A. *et al.* Developing the injury prevention and safety promotion thesaurus, international English edition: An interdisciplinary tool for indexing and searching for research literature. Progress report 1. *Safety Science* [S.I.], v. 44, n. 4, p. 279-296, 2006.

LIAO, S.-H. Knowledge management technologies and applications--literature review from 1995 to 2002. *Expert Systems with Applications* [S.I.], v. 25, n. 2, p. 155-164, 2003.

LINDBERG, D.; HUMPHREYS, B.; MCCRAY, A. The Unified Medical Language System. *Methods of information in Medicine* [S.I.], v. 32, n. 4, p. 281, 1993.

LOVELL, N.; CELLER, B. Information technology in primary health care. *International journal of medical informatics* [S.I.], v. 55, n. 1, p. 9-22, 1999.

LUCENE, A. Apache lucene. 2005.

LUDL, H.; SCHÖPE, K.; MANGELSDORF, I. Searching for information on toxicological data of chemical substances in selected bibliographic databases -- Selection of essential databases for toxicological researches. *Chemosphere* [S.I.], v. 32, n. 5, p. 867-880, 1996.

LÜLLMANN, H.; LANGELOH, A. *Farmacologia: texto e atlas*. Artmed, 2008.

LUNDSGAARDE, H.; MORESHEAD, G. Evaluation of a Computerized Clinical Information System (Micromedex). American Medical Informatics Association, 1991. p.18.

LUSHBOUGH, C.; BERGMAN, M. K.; LAWRENCE, C. J. *et al.* BioExtract Server—An Integrated Workflow-Enabling System to Access and Analyze Heterogeneous, Distributed Biomolecular Data. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* [S.I.], v. 7, n. 1, p. 12-24, 2010.

MANICA, H.; DANTAS, M.; TODESCO, J. Ontologia para Compartilhamento e Representação de Conhecimento em Saúde. *Revista Diálogos & Saberes* [S.I.], v. 4, n. 1, 2009.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. *An introduction to information retrieval*. Cambridge University Press New York, NY, USA, 2008.

MATTINGLY, C.; COLBY, G.; FORREST, J. *et al.* The Comparative Toxicogenomics Database (CTD). *Environmental Health Perspectives* [S.I.], v. 111, n. 6, p. 793, 2003.

MCBRIDE, B. Jena: A semantic web toolkit. *IEEE Internet Computing* [S.I.], p. 55-59, 2002.

MICROMEDEX, T. Micromedex healthcare series. 2010. Disponível em:<http://www.micromedex.com/>. Acesso em: 26/10/2010.

MILLER, P.; BLACK, H. HT-ATTENDING. *Journal of Medical Systems* [S.I.], v. 8, n. 3, p. 181-187, 1984.

MILLER, R.; GEISSBUHLER, A. Clinical diagnostic decision support systems–An overview. *Clinical decision support systems: Theory and practice* [S.I.], p. 3-34, 1999.

MILLER, R.; POPLE JR, H.; MYERS, J. INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine* [S.I.], v. 307, n. 8, p. 468-476, 1982.

MIN, H.; MANION, F. J.; GORALCZYK, E. *et al.* Integration of prostate cancer clinical data using an ontology. *Journal of Biomedical Informatics* [S.I.], v. 42, n. 6, p. 1035-1045, 2009.

MINACK, E.; SAUERMANN, L.; GRIMNES, G. *et al.* The Sesame LuceneSail: RDF queries with full-text search. *NEPOMUK Consortium, Technical Report* [S.I.], v. 1, 2008.

MORIK, K. Underlying assumptions of knowledge acquisition and machine learning. *Knowledge Acquisition* [S.I.], v. 3, n. 2, p. 137-156, 1991.

MUDAN, C.; DING, X.; YIJUN, L. *et al.* A Knowledge-Based Framework for Data Integration. In: Information Science and Engineering (ICISE), 2009 1st International Conference on. 2009. p.732-735. Disponível em:<10.1109/ICISE.2009.51>. Acesso em.

MUDUNURI, U.; STEPHENS, R.; BRUINING, D. *et al.* botXminer: mining biomedical literature with a new web-based application. *Nucleic Acids Res* [S.I.], v. 34, n. Web Server issue, p. W748-52, Jul 1 2006.

MUSEN, M.; VAN BEMMEL, J. *Handbook of medical informatics*. Bohn Stafleu Van Loghum, 1997.

MUSEN, M. A. An overview of knowledge acquisition*Second generation expert systems*: Springer-Verlag New York, Inc., 1993. p. 405-427.

NANDI, A.; BERNSTEIN, P. A. HAMSTER: using search clicklogs for schema and taxonomy matching. *Proc. VLDB Endow.* [S.I.], v. 2, n. 1, p. 181-192, 2009.

NIELSEN, J. Usability inspection methods. *Conference companion on Human factors in computing systems*. Boston, Massachusetts, United States: ACM, 1994. p. 413-414.

NIGEL, S. Constructing Knowledge-Based Systems. In: ENRICO, M.; ALAIN, R. (Ed.). v. 101993. p. 34-38.

NLM, N. L. O. M.-. MeSH Tree Structures. 2006. Disponível em:<http://www.nlm.nih.gov/mesh/intro_trees2006.html>. Acesso em: 27/05/2010.

NLM, N. L. O. M. TOXNET - Toxicology Data Network. v. 2010, n. 10/25/2010, 2003. Disponível em:<http://toxnet.nlm.nih.gov/>.

NOVELLO, T. Ontologias, Sistemas baseados em conhecimento e modelos de banco de dados. *Universidade Federal do Rio Grande do Sul* [S.I.], 2002.

OLSON, D.; DELEN, D. *Advanced data mining techniques*. Springer Verlag, 2008.

ORLIKOWSKI, W.; BAROUDI, J. Studying information technology in organizations: Research approaches and assumptions. *Information systems research* [S.I.], v. 2, n. 1, p. 1-28, 1991.

PAREDES-MORENO, A.; MARTÍNEZ-LÓPEZ, F. J.; SCHWARTZ, D. G. A methodology for the semi-automatic creation of data-driven detailed business ontologies. *Information Systems* [S.I.], v. 35, n. 7, p. 758-773, 2010.

PARIS, C.; SWARTOUT, W.; MANN, W. *Natural language generation in artificial intelligence and computational linguistics*. Springer, 1991.

PASQUIER, C. Biological data integration using Semantic Web technologies. *Biochimie* [S.I.], v. 90, n. 4, p. 584-594, 2008.

PREDA, N.; SUCHANEK, F. M.; KASNECI, G. *et al.* ANGIE: active knowledge for interactive exploration. *Proc. VLDB Endow.* [S.I.], v. 2, n. 2, p. 1570-1573, 2009.

PREECE, A.; HUI, K.; GRAY, A. *et al.* The KRAFT architecture for knowledge fusion and transformation. *Knowledge-Based Systems* [S.I.], v. 13, n. 2-3, p. 113-120, 2000.

REEVE, L.; HAN, H. Survey of semantic annotation platforms. ACM, 2005. p.1638.

RIBEIRO, D. I. J.; TOURINHO, F. S. V.; SAVARIS, A. *et al.* Modeling and Creation of an Ontology to Organize Knowledge related to Toxicology. *8th International Information and Telecommunication Technologies Symposium*. v. 8. Florianópolis: Fundação Barddal de Educação e Cultura, 2009. p. 175-178.

RIOS, J. Ontologias: alternativa para a representação do conhecimento explícito organizacional. 2005.

ROBINSON, L.; MCILWAINE, I.; COPESTAKE, P. *et al.* Comparative evaluation of the performance of online databases in answering toxicology queries. *International Journal of Information Management* [S.I.], v. 20, n. 1, p. 79-87, 2000.

ROCHA, C.; SCHWABE, D.; ARAGAO, M. P. A hybrid approach for searching in the semantic web. *Proceedings of the 13th international conference on World Wide Web*. New York, NY, USA: ACM, 2004.

RTECS, R. Registry of toxic effects of chemical substances. *National Institute for Occupational Safety and Health. CD-ROM* [S.I.], 2000.

RUSSELL, S.; NORVIG, P. *Artificial intelligence: a modern approach*. Prentice hall, 2009.

SAHOO, S. S.; BODENREIDER, O.; RUTTER, J. L. *et al.* An ontology-driven semantic mashup of gene and biological pathway information: Application to the domain of nicotine dependence. *Journal of Biomedical Informatics* [S.I.], v. 41, n. 5, p. 752-765, 2008.

SALTON, G.; MCGILL, M. *Introduction to modern information retrieval*. McGraw-Hill New York, 1983.

SAUNDERS, M.; LEWIS, P.; THORNHILL, A. *Research Methods for Business Students*. Financial Times Prent.Int, 2002.

SCHREIBER, A.; BIRMINGHAM, W. The sisyphus-vt initiative. *International Journal of Human-Computer Studies* [S.I.], v. 44, n. 3/4, 1996.

SCHREIBER, G. *Knowledge engineering and management: the CommonKADS methodology*. the MIT Press, 2000.

SCHREIBER, G.; WIELINGA, B.; AKKERMANS, H. *et al.* CML: The commonKADS conceptual modelling language. In: STEELS, L. *et al* (Ed.). *A Future for Knowledge Acquisition*: Springer Berlin / Heidelberg, 1994. p. 1-25. (Lecture Notes in Computer Science).

SCHREIBER, G.; WIELINGA, B.; JANSWEIJER, W. The KACTUS view on the'O'word. Citeseer, 1995. p.159-168.

SCOTNEY, B.; MCCLEAN, S. Efficient knowledge discovery through the integration of heterogeneous data. *Information and Software Technology* [S.I.], v. 41, n. 9, p. 569-578, 1999.

SCOTNEY, B.; MCCLEAN, S. Knowledge discovery from databases on the semantic Web. In: Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on. 2004. p.333-336. Disponível em:<10.1109/SSDM.2004.1311225>. Acesso em.

SENG, J.-L.; KONG, I. L. A schema and ontology-aided intelligent information integration. *Expert Systems with Applications* [S.I.], v. 36, n. 7, p. 10538-10550, 2009.

SEWELL, W.; BEVAN, A. Nonmediated use of MEDLINE and TOXLINE by pathologists and pharmacists. *Bull Med Libr Assoc* [S.I.], v. 64, n. 4, p. 382-91, Oct 1976.

SHAW, M.; GAINES, B. The synthesis of knowledge engineering and software engineering, 1992. p. 208-220.

SHORTLIFFE, E. Computer-based medical consultations: MYCIN. 1976.

SIM, I.; GORMAN, P.; GREENES, R. A. *et al.* Clinical Decision Support Systems for the Practice of Evidence-based Medicine. *Journal of the American Medical Informatics Association* [S.I.], v. 8:, p. 527-534, November 1, 2001; 2001.

SODERGREN, L. MEDLARS II: a review. *Bull Med Libr Assoc* [S.I.], v. 61, n. 4, p. 400-7, Oct 1973.

SOLR, A. Welcome to Solr. v. 2009, n. 31/02/2009, 2007. Disponível em:<http://lucene.apache.org/solr/>.

SOWA, J. Building, sharing, and merging ontologies. *Tutorial.[S. 1.: sn]* [S.I.], 1999.

STEELS, L. The componential framework and its role in reusability*Second generation expert systems*: Springer-Verlag New York, Inc., 1993. p. 273-298.

STEWART, T. A. *Intellectual capital: the new wealth of organizations*. Doubleday, 1997.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, v. 25, n. 1-2, p.161-197, 1998. Disponível em:<http://www.sciencedirect.com/science/article/B6TYX-3SYXJ6S-G/2/67ea511f5600d90a74999a9fef47ac98>. Acesso em: 01 ago 2010.

SUAREZ-ALMAZOR, M. E.; BELSECK, E.; HOMIK, J. *et al.* Identifying Clinical Trials in the Medical Literature with Electronic Databases: MEDLINE Alone Is Not Enough. *Controlled Clinical Trials* [S.I.], v. 21, n. 5, p. 476-487, 2000.

SUJANSKY, W. Heterogeneous Database Integration in Biomedicine. *Journal of Biomedical Informatics* [S.I.], v. 34, n. 4, p. 285-298, 2001.

SUN, X.; ZHU, H.; GU, J. *et al.* Research on the Semantic Web-Based Technology of Knowledge Integration for Agricultural Production. In: Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on. 2009. p.361-366. Disponível em:<10.1109/FSKD.2009.58>. Acesso em.

SWARTOUT, B.; GIL, Y. EXPECT: Explicit representations for flexible acquisition. Citeseer, 1995.

TANG, J.; ARNI, T.; SANDERSON, M. *et al.* Building a diversity featured search system by fusing existing tools. *Evaluating Systems for Multilingual and Multimodal Information Access* [S.I.], p. 560-567.

THEOBALD, M.; BAST, H.; MAJUMDAR, D. *et al.* TopX: efficient and versatile top-k query processing for semistructured data. *The VLDB Journal* [S.I.], v. 17, n. 1, p. 81-115, 2008.

TSOUMAKAS, G.; ANGELIS, L.; VLAHAVAS, I. Clustering classifiers for knowledge discovery from physically distributed databases. *Data & Knowledge Engineering* [S.I.], v. 49, n. 3, p. 223-242, 2004.

TZITZIKAS, Y.; SPYRATOS, N.; CONSTANTOPOULOS, P. Mediators over taxonomy-based information sources. *The VLDB Journal* [S.I.], v. 14, n. 1, p. 112-136, 2005.

USCHOLD, M.; GRUNINGER, M. Ontologies: Principles, methods and applications. *To appear in Knowledge Engineering Review* [S.I.], v. 11, n. 2, 1996.

USCHOLD, M.; KING, M. *Towards a methodology for building ontologies*. Citeseer, 1995.

VAN DER SPEK, R.; SPIJKERVET, A. Knowledge management: dealing intelligently with knowledge. *Knowledge management and its integrative elements* [S.I.], p. 31-59, 1997.

VAN RIJSBERGEN, C. Foundation of evaluation. *Journal of Documentation* [S.I.], v. 30, n. 4, p. 365-373, 1974.

VAN RIJSBERGEN, C. A non-classical logic for information retrieval. *The computer journal* [S.I.], v. 29, n. 6, p. 481, 1986.

VILLANUEVA-ROSALES, N.; DUMONTIER, M. yOWL: An ontology-driven knowledge base for yeast biologists. *Journal of Biomedical Informatics* [S.I.], v. 41, n. 5, p. 779-789, 2008.

WANG, Y.-H.; JHUO, P.-S. A Semantic Faceted Search with Rule-based Inference *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009* [S.I.], v. Vol I, 2009.

WASSOM, J. Use of selected toxicology information resources in assessing relationships between chemical structure and biological activity. *Environmental Health Perspectives* [S.I.], v. 61, p. 287, 1985.

WEXLER, P. TOXNET: An evolving web resource for toxicology and environmental health information. *Toxicology* [S.I.], v. 157, n. 1-2, p. 3-10, 2001.

WHITE, R. Implicit feedback for interactive information retrieval. Citeseer, 2005. p.70-70.

WIELINGA, B.; SANDBERG, J.; SCHREIBER, G. Methods and techniques for knowledge management: What has knowledge engineering to offer? *Expert Systems with Applications* [S.I.], v. 13, n. 1, p. 73-84, 1997.

WIELINGA, B. J.; SCHREIBER, A. T.; BREUKER, J. A. KADS: a modelling approach to knowledge engineering. *Knowledge Acquisition* [S.I.], v. 4, n. 1, p. 5-53, 1992.

WILCZYNSKI, N. L.; HAYNES, R. B. EMBASE search strategies for identifying methodologically sound diagnostic studies for use by clinicians and researchers. *BMC Med* [S.I.], v. 3, p. 7, 2005.

WORBOYS, M. F.; DEEN, S. M. Semantic heterogeneity in distributed geographic databases. *SIGMOD Rec.* [S.I.], v. 20, n. 4, p. 30-34, 1991.

WRIGHT, L. L. Searching fee and non-fee toxicology information resources: an overview of selected databases. *Toxicology* [S.I.], v. 157, n. 1-2, p. 89-110, 2001.

XIAOHUA, H.; WU, D. Data Mining and Predictive Modeling of Biomolecular Network from Biomedical Literature Databases. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* [S.I.], v. 4, n. 2, p. 251-263, 2007.

YE, F.; DING, X. Manufacturing enterprise business process ontology modeling for knowledge integration. In: Grey Systems and Intelligent Services, 2009. GSIS 2009. IEEE International Conference on. 2009. p.1365-1369. Disponível em:<10.1109/GSIS.2009.5408125>. Acesso em.

YOO, S.; CHOI, J. On the Query Reformulation Techniquefor Effective MEDLINE Document Retrieval. *Journal of Biomedical Informatics* [S.I.], 2010.

YU, J. Domain-oriented knowledge integration model in distributed environment. In: Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on. 2010. p.1-5. Disponível em:<10.1109/ICIME.2010.5478196>. Acesso em.

APÊNDICE A – Revisão Sistemática da Literatura – TI em apoio à Toxicologia

Esta tabela discrimina e ilustra todos os trabalhos explorados na revisão sistemática da literatura no contexto de TI em apoio à Toxicologia, fazendo uma relação entre os nomes das iniciativas nesta área e trabalhos que utilizaram estas abordagens em suas publicações.

(Lindberg *et al.*, 1993)		**UMLS**	**MBC**	**Medicina**	**Unified Medical Language System. Metatesauros de conexão conceitual que contém tesauros, bases de conhecimento/dados e vocabulários da área médica.**
(Wexler, 2001; Nlm, 2003)	**+**	**TOXNET**	**MBC**	**Toxicologia, produtos químicos perigosos, ambiente**	**Sistema de bancos de dados referentes às áreas especificadas disponíveis para acesso via WEB, desenvolvido pela National Library of Medicine da NIH.**
(Fonger *et al.*, 2000; Wexler, 2001; Nlm, 2003)	I I I I I	HSDB	BC	Produtos químicos	**H**azardous **S**ubstances **D**ata **B**ank – Banco de dados sobre produtos químicos possuindo 150 atributos relacionados a estas substâncias.

	↳				
(Wexler, 2001; Nlm, 2003)	I I I ↳	CCRIS	BC	Carcinogênese química	Banco de dados contido no TOXNET e possui aproximadamente 8000 registros sobre substâncias químicas relacionadas cientificamente avaliadas e referenciadas.
(Wassom, 1985; Wexler, 2001; Nlm, 2003)	I I I ↳	GENE-TOX	BC	Mutagênese	Possui um número aproximado de 3000 registros químicos contendo informações sobre toxicologia genética avaliados por pontos através de literatura científica.
(Sewell e Bevan, 1976; Wassom, 1985; Berman *et al.*, 1992; Ludl *et al.*, 1996; Gehanno *et al.*, 1998; Anderson *et al.*, 2000; Robinson *et*	**I I I I I I I I I I I**	**TOXLINE**	**MBC**	**Efeitos Bioquímicos, farmacológicos, fisiológicos e toxicológicos de agentes químicos.**	***Mashup* de bancos de dados que contém informações sobre toxicologia das bases literárias contidas na NLM.**

***al.*, 2000; Wexler, 2001; Wright, 2001; Nlm, 2003)**	**I I I I I I I I ↳ +**				
(Wassom, 1985; Wexler, 2001; Nlm, 2003)	I I I ↳	EMIC	BC	Testes para atividade genotóxica com agentes químicos, biológicos e físicos	Bancos de dados referentes a áreas específicas da toxicologia.
(Wassom, 1985; Wexler, 2001; Nlm, 2003)	I I I ↳	DART / ETIC	BC	Toxicologia reprodutiva	
(Wexler, 2001; Nlm, 2003)	I I I I I	TRI	BC	Controle Ambiental	Toxics Release Inventory – banco de dados alimentado pela agência de proteção ambiental (EPA) lista quantidades de produtos químicos lançados no ar, água, solo, etc., e

	↳				montantes transferidos para estações de tratamento específicas.
(Berman *et al.*, 1992; Wexler, 2001; Nlm, 2003)	I I I ↳	ChemIdPlus	BC	Identificador de produtos químicos	Trata-se de um desambiguador para identificação de produtos químicos para a determinação de sinônimos e registro do código CAS (Chemical Abstract Service).
(Lawrence *et al.*, 2006)	↳	IPSP thesaurus	CIS /BD	Segurança da saúde	Ferramenta para indexação e recuperação / tesauros para prevenção de lesões e promoção de segurança
(Ludl *et al.*, 1996; Gehanno *et al.*, 1998; Anderson *et al.*, 2000; Robinson *et al.*, 2000; Wright, 2001; Alpi, 2005)	↳	BIOSIS	BD		
(Mattingly *et al.*, 2003)	↳	CTD	BD	Química, genética, proteínas e suas relações.	Banco de dados que contem informações sobre quimica, genética, proteína e suas interações com vertebrados e

					invertebrados.
(Lundsgaarde e Moreshead, 1991; Wright, 2001; Cimino *et al.*, 2003; Micromedex, 2010)	↳ +	**MICROMEDEX**	**CIS/MDB**	**Apoio à decisão clínica**	**Projeto iniciado em 1987 com o intuito de desenvolver um software de apoio à decisão clínica.**
(Lundsgaarde e Moreshead, 1991; Micromedex, 2010)	I I ↳	POSINDEX	BD	Substâncias venenosas	Sistema de identificação e manipulação de substâncias venenosas
(Lundsgaarde e Moreshead, 1991; Micromedex, 2010)	I I ↳	DRUGDEX	BD	Agentes tóxicos	Guia de referência sobre agentes tóxicos e terapêuticos
(Lundsgaarde e Moreshead, 1991; Micromedex, 2010)	I I ↳	EMERGINDEX	BD	Doenças agudas	Ferramenta de referência para doenças agudas/médica/cirúrgica e lesões traumáticas

(Lundsgaarde e Moreshead, 1991; Micromedex, 2010)	I I ↳	IDENTIDEX	BD	Identificação de medicamentos	Identificador de medicamentos (capsula/tablete)
(Sodergren, 1973; Lindberg *et al.*, 1993; Robinson *et al.*, 2000; Wright, 2001)	↳	MEDLARS	BD	Publicações/medicina	Medical Literature Analysis and retrieval System. Trata-se de um sistema de análise e recuperação de publicações da área de medicina.
(Anderson *et al.*, 2000; Robinson *et al.*, 2000)	↳	TRACE	BD	Toxicologia	Specialist bibliographic database. Inclui detalhes de documentos sobre este contexto inclusos na base desde 1987, contendo mais de 12000 registros.
(Bateman *et al.*, 2002)	↳	TOXBASE	BD/ABBC	Substâncias venenosas	Banco de dados sobre substâncias venenosas, utilizado no Reino Unido.
(Alpi, 2005; Nlm, 2006)	↳	MeSH	BD/ABBC	Médica	Vocabulário controlado utilizado para indexação de artigos nos portais PubMED e MEDLINE
(Sewell e Bevan, 1976;	↳ +	MEDLINE	BD/ABBC	Publicações/medicina	Portal para indexação de artigos científicos relativos à área da

Kenton e Scott, 1978; Lindberg *et al.*, 1993; Ludl *et al.*, 1996; Gehanno *et al.*, 1998; Anderson *et al.*, 2000; Robinson *et al.*, 2000; Suarez-Almazor *et al.*, 2000; Cimino *et al.*, 2003; Alako *et al.*, 2005; Alpi, 2005; Darmoni *et al.*, 2006; Mudunuri *et al.*, 2006)					saúde
(Mudunuri *et al.*, 2006)	I ↳	botXminer	(CIS/CDSS/ SE)	Mineração em BD	Aplicação que realiza buscas em arquivos XML fornecidos por MEDLINE

(Alako *et al.*, 2005)	I | ↳	CoPub Mapper	(CIS/CDSS/ SE)	Mineração em BD	Aplicação que realiza buscas na área de genética/co-ocorrência genética em arquivos XML fornecidos por MEDLINE
(Darmoni *et al.*, 2006)	I ↳	MCA	(CIS/CDSS/ SE)	Algoritmo	MEDLINE categorization algorithm. Algoritmo de categorização de conteúdo MEDLINE
(Ludl *et al.*, 1996; Gehanno *et al.*, 1998; Anderson *et al.*, 2000; Robinson *et al.*, 2000; Suarez-Almazor *et al.*, 2000; Wright, 2001; Alpi, 2005; Wilczynski e Haynes, 2005)	↳	EMBASE	BD/ABBC	Publicações/medicina	Base de dados contendo informações sobre dados clínicos/médicos de âmbito geral.
(Huismans, 1980;	↳	IRPTC	BD	Substâncias venenosas/agentes	*Storehouse* que contém informações sobre substâncias

Kurlyandskiy e Sidorov, 2003)				tóxicos	perigosas e associações com agentes tóxicos
(Rtecs, 2000; Kurlyandskiy e Sidorov, 2003)	↳	RTECS	BD	Substancias toxicas	Registry of Toxic effects of Chemical Substances
(Shortliffe, 1976)	↳	Mycin	(CIS/CDSS/ SE)	Medicina	Sistema desenvolvido na Universidade de Santford com o intuito de recomendar medicamentos para tratamento de infecção por bacterias.
(Miller e Black, 1984)	↳	HT-Attending	(CIS/CDSS/ SE)	Farmacia	Sistema computacional desenvolvido para efetuar críticas ao manuseio farmacológico relativo à hipertensão
(Miller *et al.*, 1982)	↳	INTERNIST-I	(CIS/CDSS/ SE)	Medicina	Sistema experimental com capacidade de gerar múltiplos e complexos diagnósticos na área de medicina.
(Berman *et al.*, 1992)	↳	DBX	(CIS/CDSS/ SE)	Informações biomédicas	Sistema para provisão de informações biomédicas.

APÊNDICE B – Revisão Sistemática da Literatura – Engenharia do Conhecimento para integração de bases de dados heterogêneas

Esta tabela discrimina e ilustra todos os trabalhos explorados na revisão sistemática da literatura no contexto de Engenharia do Conhecimento para integração de bases de dados heterogêneas, descrevendo sucintamente características de cada trabalho.

referência	Mecanismo de interfaceamento	área de aplicação	Método de avaliação dos resultados	Abordagem	Indexação	Ampliação de escopo da pesquisa	OBS.:
(Bellatreche *et al.*, 2006)	Proprietário	Genérica	Discussão teórica	Integração remota	-	-	Abordagem KB/Mecanismo
(Gupta *et al.*, 2010)	Proprietário	Informações Biológicas	Avaliação autocontida	Integração por extração	ontologia	-	Abordagem KB/Mecanismo
(Villanueva-Rosales e Dumontier, 2008)	Proprietário	Informações Biológicas	Discussão teórica	Integração por extração	-	Reasoner (Protègè)	Abordagem KB. As consultas devem seguir um padrão específico
(Scotney e Mcclean, 2004)	Proprietário	Genérica	Discussão teórica	Integração remota	-	Ontologia associada	Abordagem KB/Mecanismo

referência	Mecanismo de interfaceamento	área de aplicação	Método de avaliação dos resultados	Abordagem	Indexação	Ampliação de escopo da pesquisa	OBS.:
(Min *et al.*, 2009)	Independente (utilização SPARQL)	Medicina	Ilustração dos artefatos produzidos	Integração remota	-	-	Abordagem KB. Construção de uma ontologia para desambiguação de conceitos relacionados à câncer de próstata.
(Paredes-Moreno *et al.*, 2010)	-	e-Business	Análise comparativa	Integração remota	-	-	Abordagem KB. Trata da implementação de uma ontologia que visa nortear as pesquisas utilizando anotações semânticas na área de Business.
(Alonso-Calvo *et al.*, 2007)	Independente (JADE)	Informações Biomédicas	Discussão teórica	Integração por extração		-	Abordagem KB / Mecanismo. Integração de bases de conhecimento na área de Engenharia Biomédica. Extração semi-automática via Crawlers.

referência	Mecanismo de interfaceamento	área de aplicação	Método de avaliação dos resultados	Abordagem	Indexação	Ampliação de escopo da pesquisa	OBS.:
(Sahoo *et al.*, 2008)	Independente (SPARQL)	Informações biomédicas (Domínio dependência nicotina)	Exemplificação quantitativa	Integração remota	-	Reasoner	Abordagem KB. Mashup ontológico com integração remota, aplicado ao domínio da dependência de nicotina utilizando bases de conhecimento de ciências biológicas
(Seng e Kong, 2009)	Proprietário	e-business	Ilustração dos artefatos produzidos	Integração remota	-	Reformulação de consulta	Abordagem KB / Mecanismo. Mashup de bases de dados heterogêneas auxiliado por uma ontologia para contexto e-business.
(Pasquier, 2008)	Independente (SPARQL)	Ciências biológicas	Ilustração dos artefatos produzidos	Integração por extração	-	-	Abordagem KB / Utilização de métodos simples para extração do conhecimento contido nas bases relacionadas à biologia.

referência	Mecanismo de interfaceamento	área de aplicação	Método de avaliação dos resultados	Abordagem	Indexação	Ampliação de escopo da pesquisa	OBS.:
	-	Genérica	Discussão teórica	Integração remota / mapeamento	Ontologia	-	Interface de mapeamento de metadados entre ontologias heterogêneas.
(Aijuan e Honglin, 2005)	-	Virtual Learning	Discussão teórica	Integração remota	Ontologia	-	Virtual Learning (Estudo de caso em ciências biológicas). Utilização de mapeamento de conhecimento através de ontologias.
(Ha *et al.*, 2008)	Proprietário	ERP	Discussão teórica	Integração remota	Ontologia	-	Abordagem KB / Mecanismo. Tomada de decisão em ERP no Mercado Chinês.
(Lushbough *et al.*, 2010)	Proprietário	DNA / Protein	Discussão teórica	Integração por extração	-	-	Sistema que provê acesso a bases biomoleculares através de bases de conhecimento, ferramentas de análise e workflows.

referência	Mecanismo de interfaceamento	área de aplicação	Método de avaliação dos resultados	Abordagem	Indexação	Ampliação de escopo da pesquisa	OBS.:
(Xiaohua e Wu, 2007)	Proprietário	Ciências Biológicas	Ilustração do artefato produzido em estudo de caso	Integração por extração (data mining)	-	-	Data mining de informações sobre ciências biológicas. Estudo com apelo à genética. Serve como ilustração, mas não como information retrieval
(Sun *et al.*, 2009)	Proprietário	Agricultura	Quantitativos (precision / recall)	Integração remota	Ontologia	-	Abordagem KB / Mecanismo. Mecanismo de descoberta de conhecimento em bases de dados heterogêneas no contexto de agricultura. Integração remota.

referência	Mecanismo de interfacea-mento	área de aplicação	Método de avaliação dos resultados	Aborda-gem	Indexa-ção	Ampliação de escopo da pesquisa	OBS.:
(Yu, 2010)	Proprietário	Genérico	Discussão teórica	Integração remota	Ontologia	Módulo semântico	Abordagem KB / Mecanismo. Mecanismo modular disposto na filosofia grid/modular com o intuito de integrar bases de dados heterogêneas de contexto genérico.
(Mudan *et al.*, 2009)	Proprietário	Genérico	Quantitativos	Integração por extração (Matching Semântico)	-	-	Abordagem KB / Mecanismo. Utilização de mapeamento por matching semântico para integração através da extração dos conceitos das bases de dados heterogêneas.

referência	Mecanismo de interfaceamento	área de aplicação	Método de avaliação dos resultados	Abordagem	Indexação	Ampliação de escopo da pesquisa	OBS.:
(Biffl *et al.*, 2010)	Proprietário	Documentação de projetos	Comparativa. Resultados anteriores à aplicação do método VS. Resultados atuais	Integração por extração	Ontologia	Reasoner	Abordagem KB. Utilização da integração de diversas fontes de conhecimento relativos à gerência de projetos no contexto de desenvolvimento, utilizando fontes como bug-tracker e svn.
(Ye e Ding, 2009)	-	Business process	Discussão teórica	Integração por mapeamento	Ontologia	-	Abordagem KB. Integração do conhecimento relacionado à business process. Utilização de uma ontologia de mapeamento do conhecimento.

referência	Mecanismo de interfaceamento	área de aplicação	Método de avaliação dos resultados	Abordagem	Indexação	Ampliação de escopo da pesquisa	OBS.:
(Gong *et al.*, 2005)	proprietário	Healthcare applications	Quantitativa / estatísticas	Integração remota	Ontologia		Abordagem KB / Mecanismo. Criação de um middleware para integração de sistemas HIS, RIS e PACS utilizando como mecanismo mapeador uma ontologia.
(Gubanov *et al.*, 2009)	Proprietário	Genérico	Discussão teórica	Integração remota	-	Navegação na árvore	Centralizador de uso generico, desenvolvido pela ibm. Possui uma camada de abstração de alto nível utilizando outros mecanismos de busca, através de uma emulação de mashup.

referência	Mecanismo de interfaceamento	área de aplicação	Método de avaliação dos resultados	Abordagem	Indexação	Ampliação de escopo da pesquisa	OBS.:
(Doerr e Iorizzo, 2008)	Proprietário	Genérico	Discussão teórica	Integração remota	-	-	Mecanismo de busca para contexo genérico que utiliza tecnologias de websemântica para integração e busca do conhecimento em bases dedados heterogêneas
(Tzitzikas *et al.*, 2005)	Proprietário	Genérico	Quantitativos	Integração remota	Taxonomia	Query translator module	Abordagem KB / Mecanismo. Proposta de um "mediador" para integração de taxonomias de dados, através de uma taxonomia de abstração.

referência	Mecanismo de interfaceamento	área de aplicação	Método de avaliação dos resultados	Abordagem	Indexação	Ampliação de escopo da pesquisa	OBS.:
(Preda *et al.*, 2009)	Independente (SPARQL)	Genérico	Ilustração dos artefatos produzidos	Integração remota	-	Query translator module	Abordagem KB / Mecanismo. Trata-se de um mecanismo de QA que possui uma base local para respostas. Quando a resposta não é encontrada no banco, automaticamente são utilizados crawlers para buscar este tipo de conhecimento na web.
(Theobald *et al.*, 2008)	Proprietário	Genérico	Quantitativos - estatísticos	Integração remota	Índices invertidos	Query expansion	Abordagem KB / Mecanismo. Uso genérico. Integração remota de documentos semi-estruturados.utilizando XPATH. Suporta pesquisas em dados semi-estruturadas e em texto livre.

referência	Mecanismo de interfaceamento	área de aplicação	Método de avaliação dos resultados	Abordagem	Indexação	Ampliação de escopo da pesquisa	OBS.:
(Koufi *et al.*, 2010)	Proprietário	Saúde	Discussão teórica	Integração remota	-	-	Abordagem Mecanismo. Portal que visa prover conhecimento necessário para apoio a diagnostico e tratamentos efetuados na residência do paciente.
(Nandi e Bernstein, 2009)	Proprietário	Data warehouses	Quantitativos (estatísticos: precision x recall)	Integração remota	Taxonomia	Clicklogs	Abordagem KB / Mecanismo. Utilização de técnicas chamadas clicklogs (log para registrar o comportamento do usuário) para matching em data warehouses.
(Eyal e Milo, 2001)	Proprietário	e-business	Ilustração dos artefatos produzidos	Integração remota	-	-	Abordagem KB / Mecanismo. Integração de conteúdo e-business utilizando wrappers para obtenção de conteúdo XML.

referência	Mecanismo de interfaceamento	área de aplicação	Método de avaliação dos resultados	Abordagem	Indexação	Ampliação de escopo da pesquisa	OBS.:
(Bonifati *et al.*, 2010)	Proprietário	P2P applications	Quantitativos (autocontidos)	Integração remota	-	Query reformulation	Abordagem KB / Mecanismo. Aplicação para integração de bases heterogêneas em aplicações P2P com utilização de query reformulation como mecanismo auxiliador para maior abrangência da pesquisa.

APÊNDICE C – Totalização para os resultados (IR, QE) utilizando as métricas Average Precision e P@10

Tabela 9 - Tabela com a totalização para os resultados com os métodos de Recuperação de Informação (IR) e o comparativo com a inserção da técnica de Expansão de Consulta (QE) utilizando as métricas Average Precision e P@10

Tópico	#	IR (average Precision)	IR (P@10)	IR + QE (average Precision)	IR + QE (P@10)
Doxycycline	1	0,470	0,40	0,470	0,400
Oxcarbazepine	2	0,057	0,20	0,057	0,200
Cephalexin	3	0,071	0,40	0,039	0,300
Verapamil hydrochloride	4	0,068	0,10	0,068	0,100
Hydrochlorothiazide	5	0,060	0,40	0,060	0,400
Phenobarbital	6	0,222	0,50	0,174	0,500
Betamethasone	7	0,111	0,10	0,111	0,100
Cefazolin	8	0,039	0,10	0,039	0,100
Gentamicin	9	0,042	0,30	0,042	0,300
Propranolol hydrochloride	10	0,322	0,10	0,322	0,100
Ampicillin	11	0,060	0,40	0,600	0,400
Furosemida	12	0,038	0,10	0,038	0,100
Aspirin	13	0,138	0,40	0,190	0,200
Sertraline	14	0,131	0,40	0,131	0,400
Ranitidine	15	0,045	0,10	0,045	0,100
Albendazole	16	0,071	0,10	0,071	0,100

Tópico	#	IR (average Precision)	IR (P@10)	IR + QE (average Precision)	IR + QE (P@10)
Azithromycin	17	0,053	0,30	0,053	0,300
Atenolol	18	0,066	0,40	0,066	0,400
Phenytoin	19	0,205	0,60	0,223	0,700
Carbamazepine	20	0,266	0,60	0,266	0,600
Amitriptyline hydrochloride	21	0,062	0,10	0,062	0,100
Haloperidol	22	0,116	0,20	0,116	0,200
Chlorpromazine	23	0,141	0,30	0,141	0,300
Clonazepam	24	0,094	0,40	0,094	0,400
Diazepam	25	0,106	0,30	0,104	0,300
Metformin hydrochloride	26	0,125	0,10	0,125	0,100
Chlorpropamide	27	0,266	0,40	0,266	0,400
Lithium carbonate	28	0,072	0,10	0,072	0,100
Morphine	29	0,233	0,50	0,264	0,700
Captopril	30	0,119	0,40	0,119	0,400
Meperidine	31	0,132	0,80	0,077	0,600
Acetaminophen	32	0,089	0,20	0,100	0,200
Diltiazem	33	0,102	0,40	0,102	0,400
Diclofenac	34	0,071	0,10	0,071	0,100
Ibuprofen	35	0,025	0,10	0,025	0,100
Amoxicillin	36	0,034	0,20	0,034	0,200

Tópico	#	IR (average Precision)	IR (P@10)	IR + QE (average Precision)	IR + QE (P@10)
Dexamethasone	37	0,190	0,30	0,190	0,300
Ketoconazole	38	0,071	0,10	0,071	0,100
Hydrocortisone	39	0,170	0,20	0,170	0,200
Enalapril maleate	40	0,037	0,10	0,037	0,100
Imipramine	41	0,240	0,60	0,243	0,600
Naproxen	42	0,071	0,10	0,071	0,100
Midazolam hydrochloride	43	0,029	0,10	0,029	0,100
Losartan potassium	44	0,037	0,10	0,037	0,100
Valproic acid	45	0,175	0,70	0,175	0,700
Prednisone	46	0,166	0,20	0,166	0,200
Risperidone	47	0,058	0,10	0,058	0,100
Chlorthalidone	48	0,049	0,30	0,049	0,300
Alprazolam	49	0,049	0,20	0,049	0,200
Paracetamol	50	0,089	0,20	0,100	0,200
Média		0,115	0,28	0,126	0,274
Desvio Padrão		0,088	0,18	0,113	0,184
Erro		0,012	0,026	0,016	0,026

APÊNDICE D – Totalização para os resultados (IR, SI) utilizando as métricas Average Precision e P@10

Tabela 10 - Tabela com a totalização para os resultados com os métodos de Recuperação de Informação (IR) com Aperfeiçoamento Semântico (SI) e o comparativo com a inserção da técnica de Expansão de Consulta (QE) utilizando as métricas *Average Precision* e P@10

Tópico	#	IR+SI (average Precision)	IR + SI (P@10)	IR + SI + QE (average Precision)	IR + SI + QE (P@10)
Doxycycline	1	0,570	0,700	0,570	0,700
Oxcarbazepine	2	0,227	0,700	0,227	0,700
Cephalexin	3	0,022	0,100	0,051	0,400
Verapamil hydrochloride	4	0,062	0,100	0,065	0,100
Hydrochlorothiazide	5	0,010	0,100	0,010	0,100
Phenobarbital	6	0,818	0,900	0,769	0,800
Betamethasone	7	1,000	0,900	1,000	0,900
Cefazolin	8	0,053	0,300	0,685	0,800
Gentamicin	9	0,068	0,400	0,108	0,500
Propranolol hydrochloride	10	0,000	0,000	0,000	0,000
Ampicillin	11	0,065	0,400	0,065	0,400
Furosemida	12	0,019	0,100	0,038	0,100
Aspirin	13	0,071	0,100	0,073	0,200
Sertraline	14	0,782	0,900	0,782	0,900
Ranitidine	15	0,990	1,000	0,990	1,000
Albendazole	16	0,071	0,100	0,071	0,100

Tópico	#	IR+SI (average Precision)	IR + SI (P@10)	IR + SI + QE (average Precision)	IR + SI + QE (P@10)
Azithromycin	17	0,068	0,400	0,068	0,400
Atenolol	18	0,000	0,000	0,000	0,000
Phenytoin	19	0,291	0,900	0,382	0,800
Carbamazepine	20	0,227	0,700	0,227	0,700
Amitriptyline hydrochloride	21	0,782	0,900	0,782	0,900
Haloperidol	22	0,996	1,000	0,996	1,000
Chlorpromazine	23	0,996	1,000	0,996	1,000
Clonazepam	24	0,238	0,700	0,238	0,700
Diazepam	25	0,238	0,700	0,258	0,700
Metformin hydrochloride	26	0,475	0,400	0,475	0,400
Chlorpropamide	27	0,345	0,400	0,345	0,400
Lithium carbonate	28	0,782	0,900	0,782	0,900
Morphine	29	0,282	0,800	0,323	0,800
Captopril	30	0,000	0,000	0,000	0,000
Meperidine	31	0,231	0,700	0,235	0,800
Acetaminophen	32	0,071	0,200	0,145	0,300
Diltiazem	33	0,511	0,900	0,511	0,900
Diclofenac	34	0,071	0,100	0,119	0,200
Ibuprofen	35	0,652	0,600	0,652	0,600
Amoxicillin	36	0,064	0,400	0,064	0,400

Tópico	#	IR+SI (average Precision)	IR + SI (P@10)	IR + SI + QE (average Precision)	IR + SI + QE (P@10)
Dexamethasone	37	0,960	0,900	0,960	0,900
Ketoconazole	38	0,071	0,100	0,071	0,100
Hydrocortisone	39	1,000	1,000	1,000	1,000
Enalapril maleate	40	0,498	0,900	0,498	0,900
Imipramine	41	0,782	0,900	0,782	0,900
Naproxen	42	0,071	0,100	0,071	0,100
Midazolam hydrochloride	43	0,241	0,700	0,241	0,700
Losartan potassium	44	0,498	0,900	0,498	0,900
Valproic acid	45	0,245	0,700	0,245	0,700
Prednisone	46	0,966	0,900	0,966	0,900
Risperidone	47	0,937	1,000	0,937	1,000
Chlorthalidone	48	0,010	0,100	0,010	0,100
Alprazolam	49	0,328	0,700	0,212	0,600
Paracetamol	50	0,071	0,200	0,145	0,300
Média		0,377	0,552	0,395	0,574
Desvio Padrão		0,358	0,354	0,352	0,336
Erro		0,051	0,050	0,050	0,047

APÊNDICE E – Totalização para os resultados (IR, QE) utilizando as métricas *Precision* e *Recall*

Tabela 11 - Tabela com a totalização para os resultados com os métodos de Recuperação de Informação (IR) e o comparativo com a inserção da técnica de Expansão de Consulta (QE) utilizando as métricas Precision e Recall

Topic	#	IR (precision)	IR (recall)	IR + QE (precision)	IR + QE (recall)
Doxycycline	1	0,071	0,400	0,071	0,400
Oxcarbazepine	2	0,068	0,666	0,068	0,666
Cephalexin	3	0,888	0,800	0,888	0,800
Verapamil hydrochloride	4	0,062	0,500	0,062	0,500
Hydrochlorothiazide	5	0,083	0,500	0,083	0,500
Phenobarbital	6	0,404	0,314	0,404	0,293
Betamethasone	7	0,111	0,500	0,111	0,500
Cefazolin	8	0,047	0,666	0,047	0,666
Gentamicin	9	0,069	0,272	0,069	0,272
Propranolol hydrochloride	10	0,032	0,500	0,032	0,500
Ampicillin	11	0,085	0,571	0,085	0,571
Furosemida	12	0,038	1,000	0,038	1,000
Aspirin	13	0,357	0,138	0,428	0,142
Sertraline	14	0,250	0,444	0,250	0,444
Ranitidine	15	0,045	0,083	0,045	0,083
Albendazole	16	0,071	0,250	0,071	0,250
Azithromycin	17	0,066	0,750	0,066	0,750

Topic	#	IR (precision)	IR (recall)	IR + QE (precision)	IR + QE (recall)
Atenolol	18	0,100	0,571	0,100	0,571
Phenytoin	19	0,348	0,357	0,372	0,355
Carbamazepine	20	0,413	0,444	0,413	0,444
Amitriptyline hydrochloride	21	0,062	0,500	0,062	0,500
Haloperidol	22	0,250	0,116	0,250	0,116
Chlorpromazine	23	0,187	0,136	0,187	0,136
Clonazepam	24	0,117	0,800	0,117	0,800
Diazepam	25	0,617	0,062	0,647	0,065
Metformin hydrochloride	26	0,125	0,500	0,125	0,500
Chlorpropamide	27	0,363	0,571	0,363	0,571
Lithium carbonate	28	0,125	0,111	0,125	0,111
Morphine	29	0,400	0,461	0,400	0,461
Captopril	30	0,148	0,500	0,148	0,500
Meperidine	31	0,181	0,571	0,204	0,600
Acetaminophen	32	0,113	0,238	0,136	0,230
Diltiazem	33	0,153	0,400	0,153	0,400
Diclofenac	34	0,071	0,500	0,071	0,500
Ibuprofen	35	0,025	0,100	0,025	0,100
Amoxicillin	36	0,041	0,250	0,041	0,250
Dexamethasone	37	0,300	0,166	0,300	0,166
Ketoconazole	38	0,071	0,055	0,071	0,055

Topic	#	IR (precision)	IR (recall)	IR + QE (precision)	IR + QE (recall)
Hydrocortisone	39	0,400	0,285	0,400	0,285
Enalapril maleate	40	0,037	0,500	0,037	0,500
Imipramine	41	0,375	0,461	0,375	0,461
Naproxen	42	0,071	0,250	0,071	0,250
Midazolam hydrochloride	43	0,029	0,500	0,029	0,500
Losartan potassium	44	0,037	0,500	0,037	0,500
Valproic acid	45	0,275	0,571	0,275	0,571
Prednisone	46	0,200	0,166	0,200	0,166
Risperidone	47	0,058	0,166	0,058	0,166
Chlorthalidone	48	0,061	0,750	0,061	0,750
Alprazolam	49	0,058	0,500	0,058	0,500
Paracetamol	50	0,113	0,238	0,136	0,230
	Média	0,173	0,413	0,177	0,413
	Desv.Padrão	0,173	0,221	0,176	0,221
	Erro	0,024	0,031	0,025	0,031

APÊNDICE E – Totalização para os resultados (IR, SI, QE) utilizando as métricas *Precision* e *Recall*

Tabela 12 - Tabela com a totalização para os resultados com os métodos de Recuperação de Informação (IR) com Aperfeiçoamento Semântico (SI) e o comparativo com a inserção da técnica de Expansão de Consulta (QE) utilizando as métricas Precision e Recall

Topic	#	IR + SI (precision)	IR +SI (recall)	IR + QE + SI (prec.)	IR + QE + SI (recall)
Doxycycline	1	1,000	0,049	1,000	0,049
Oxcarbazepine	2	0,448	0,270	0,448	0,270
Cephalexin	3	1,000	0,041	1,000	0,041
Verapamil hydrochloride	4	0,062	0,016	0,062	0,016
Hydrochlorothiazide	5	0,200	0,500	0,200	0,500
Phenobarbital	6	1,000	0,037	1,000	0,037
Betamethasone	7	1,000	0,152	1,000	0,152
Cefazolin	8	1,000	0,038	1,000	0,038
Gentamicin	9	0,093	0,666	0,162	0,411
Propranolol hydrochloride	10	0,032	0,083	0,032	0,083
Ampicillin	11	0,085	0,666	0,085	0,666
Furosemida	12	0,038	1,000	0,038	1,000
Aspirin	13	0,071	0,500	0,142	0,200
Sertraline	14	0,937	0,365	0,937	0,365
Ranitidine	15	1,000	0,285	1,000	0,285
Albendazole	16	0,071	1,000	0,071	1,000

Topic	#	IR + SI (precision)	IR +SI (recall)	IR + QE + SI (prec.)	IR + QE + SI (recall)
Azithromycin	17	0,888	0,666	0,888	0,666
Atenolol	18	0,100	0,571	0,100	0,571
Phenytoin	19	0,395	0,377	0,604	0,325
Carbamazepine	20	0,448	0,270	0,448	0,270
Amitriptyline hydrochloride	21	0,937	0,365	0,937	0,365
Haloperidol	22	1,000	0,210	1,000	0,210
Chlorpromazine	23	1,000	0,210	1,000	0,210
Clonazepam	24	0,411	0,291	0,411	0,291
Diazepam	25	0,411	0,291	0,500	0,073
Metformin hydrochloride	26	0,500	0,800	0,500	0,800
Chlorpropamide	27	0,363	0,800	0,363	0,800
Lithium carbonate	28	0,937	0,365	0,937	0,365
Morphine	29	0,333	0,789	0,377	0,809
Captopril	30	0,000	0,000	0,000	0,000
Meperidine	31	0,318	0,736	0,318	0,636
Acetaminophen	32	0,022	0,500	0,090	0,307
Diltiazem	33	0,615	0,271	0,615	0,271
Diclofenac	34	0,071	0,500	0,142	0,500
Ibuprofen	35	0,307	0,631	0,307	0,631
Amoxicillin	36	0,083	0,666	0,083	0,666
Dexamethasone	37	1,000	0,169	1,000	0,169

Topic	#	IR + SI (precision)	IR +SI (recall)	IR + QE + SI (prec.)	IR + QE + SI (recall)
Ketoconazole	38	0,071	1,000	0,071	1,000
Hydrocortisone	39	1,000	0,169	1,000	0,169
Enalapril maleate	40	0,592	0,271	0,592	0,271
Imipramine	41	0,937	0,365	0,937	0,365
Naproxen	42	0,071	0,500	0,071	0,500
Midazolam hydrochloride	43	0,411	0,311	0,411	0,311
Losartan potassium	44	0,592	0,271	0,592	0,271
Valproic acid	45	0,448	0,288	0,448	0,288
Prednisone	46	1,000	0,169	1,000	0,169
Risperidone	47	0,941	0,210	0,941	0,210
Chlorthalidone	48	0,020	0,500	0,020	0,500
Alprazolam	49	0,411	0,291	0,411	0,294
Paracetamol	50	0,022	0,500	0,090	0,307
	Média	0,494	0,400	0,508	0,374
	Desv.Padrão	0,386	0,268	0,378	0,267
	Erro	0,055	0,038	0,053	0,038

APÊNDICE F – Dados da análise comparativa com mecanismos semelhantes em medida de tempo

Tabela 13 - Tabela com a totalização para os resultados com os métodos tradicionalmente usados para atendimento em emergências em toxicologia clínica comparando com o método proposto computado em métrica de tempo (segundos)

Topic	#	Método tradicional **(Micromedex)**	Método tradicional #2 **(DrugDEX)**	Método proposto
Cephalexin	1	21,00	7,00	5,30
Phenobarbital	2	8,00	10,00	5,60
Sertraline	3	8,00	6,00	4,20
Albendazole	4	7,00	6,00	5,50
Carbamazepine	5	7,00	10,00	5,60
Haloperidol	6	7,00	7,00	4,20
Diazepam	7	9,00	6,00	4,80
Lithium carbonate	8	6,00	8,00	6,90
Acetaminophen	9	11,00	6,00	6,70
Imipramine	10	6,50	5,00	4,00
Média		9,05	7,10	5,28
Desvio Padrão		4,44	1,73	1,01
Erro		2,86	2,25	1,67

APÊNDICE G – Publicações

CABRAL, R. B. ; ANDRADE, Rafael ; SAVARIS, Alexandre ; ZANIN, Marlene ; Wangenheim, Aldo v. . Plataforma de Gerência do Conhecimento Aplicada em um Ambiente de Toxicologia Clínica e Toxicovigilância. In: Congresso Brasileiro de Informática na Saúde, 2008, Campos do Jordão - SP. Congresso Brasileiro de Informática na Saúde. São Paulo : SBIS, 2008.

RIBEIRO JR, D. I. ; TOURINHO, F. S. V. ; SAVARIS, Alexandre ; CABRAL, R. B. ; Wangenheim, Aldo v. . Modeling and Creation of an Ontology to Organize Knowledge related to Toxicology. In: 8th International Information and Telecommunication Technologies Symposium, 2009, Florianópolis. Proceedings of 8th International Informationa and Telecommunication Technologies Symposium. Florianópolis : Fundação Barddal de Educação e Cultura, 2009. v. 8. p. 175-178.]

CABRAL, R. B. ; ANDRADE, Rafael ; BARCELLOS JR., C. L. ; Wangenheim, Aldo v. . Semantic Information Indexing and Retrieval on Patient Medical Data. In: 8th International Information and Telecommunication Technologies Symposium, 2009, Florianópolis. Proceedings of 8th International Information and Telecommunication Technologies Symposium. Florianópolis : Fundação Barddal de Educação e Cultura, 2009. v. 8. p. 171-174.